Chapter I Measuring Visual Aesthetic Sensitivity for Everyday Objects: Scale Development and Psychometric Properties

In this chapter a general approach to scale development that is used to develop a scale for measuring aesthetic sensitivity is described. Aesthetic sensitivity is defined here as an individual's perception of the beauty of objects in his or her immediate environment. Because this research focus is on the immediate environment the stimuli used for scale development are objects taken from people's everyday environment such as cutlery, vehicles, furniture, and jewellery.

1) Theoretical Considerations

a. The Absence of Relevant Criteria

In developing a scale for measuring visual aesthetic sensitivity, which criteria can be used to choose stimuli such that they differ in their aesthetic values? Or in other words, how can stimuli be chosen such that some of them are more beautiful than others? One might think that the Gestalt Psychology contributed an answer to this question in the concept *Gute Gestalt*, "goodness of configuration." According to this concept people do not perceive their environment in terms of discrete visual elements but rather in terms of configurations (*Gestalten*). The Gestalt school formulated laws to illustrate the goodness of configuration, such as symmetry, balance, and proportion (e.g., Boring, 1942; Koffka, 1935; Metzger, 1953). However, as Berlyne (1971) pointed out, these characteristics of good configurations are rather elementary. They give general guidelines about the quality of configurations. However, these guidelines are so nonspecific that they can hardly be used as criteria to decide which of two patterns is superior aesthetically. Berlyne examined the relationship of stimulus patterns that varied on collative structural properties along dimensions like familiar-novel, simple-complex, expected-surprising etc. (Berlyne, 1963, 1970, 1974a). While this research supported the proposed relationships between the collative properties and aesthetic sensitivity, it was mostly done with artificial stimuli, and the collative properties could not be confirmed for everyday objects (Ritterfeld, 2002).

In sum, even though aesthetic principles have been discovered that might be important for aesthetic appeal of objects such as symmetry, balance, clarity, color, novelty and many others, it is not yet known for everyday objects which characteristics and thus which stimuli educe a positive response in an aesthetically sensitive perceiver. Consequently, it seems rather difficult when considering everyday objects to identify stimuli that differ in their aesthetic value. To assure that the stimuli used for the present scale development differed in their aesthetic value, stimuli were chosen that were obvious exemplars of ugly and beautiful objects. The extent to which a person agrees that the ugly stimuli are rather ugly and the beautiful stimuli are rather beautiful is then seen as indication of the amount of his or her sensitivity. That is, a person would have a high aesthetic sensitivity if he or she judged all ugly stimuli to be ugly and all beautiful stimuli to be beautiful.

b. Aesthetic Response

Although no agreement exists on how to define an aesthetic response, there is some consensus that aesthetic responses involve an affective response to the object (e.g., Bamossy et al., 1983; Cupchik, 1994; Veryzer, 1993)[1]. Holbrook and Zirlin (1985) even define aesthetic response as a "deeply felt experience that is enjoyed purely for its own sake without regard for other more practical considerations" (p. 21). As this definition illustrates, the affect is generally considered to be positive because aesthetics is by definition concerned with the beauty. However, it is also possible that the affective response to an object is negative, namely when the object is not found to be aesthetically pleasing (Veryzer, 1993). In looking at a beautiful car, for example, a person might have a positive affective response, namely in experiencing pleasure. If the person does not perceive the car as beautiful the resulting affective reaction might be neutral or negative. Following this idea, the present research aims to assess aesthetic responses that are either positive or negative affective responses to everyday objects. The objective of the aesthetic sensitivity scale is thus to measure to what extent individuals are sensitive towards the “look and feel” of an object.

[1] Even though overall judgments about everyday objects can involve functional considerations, affective responses can be considered part of the aesthetic judgment and have been investigated in the past (e.g., (Bloch, 1995; Veryzer, 1993).

c. Existing Scales

Many attempts have been made to measure visual aesthetic sensitivity. However, these measures either show poor psychometric properties (Graves, 1948; Meier, 1940; Welsh, 1949, 1987), were developed for specific experimental settings (e.g., Child, 1962, 1964, 1965; Karwoski & Christensen, 1926), are rather time-consuming (Eysenck, 1983; Götz et al., 1979), and/or focused exclusively on art works (Bamossy et al., 1983; Eysenck, 1983; Götz et al., 1979). For instance, Eysenck and others (Child, 1964; Crannell, 1953; Eysenck, 1967, 1970; Eysenck & Castle, 1971; Eysenck, Castle, Averill, Stanat, & More, 1970); (for an overview, see Eysenck, 1983; 1988) have shown that the Meier Art Judgment Test (Meier, 1940), the Maitland Graves Design Judgment Test (Graves, 1948) and the Welsh Figure Preference Test (Welsh, 1949) have poor psychometric properties.

Measures with good psychometric properties are the Visual Aesthetic Sensitivity Test (VAST, Götz et al., 1979), the Test of Aesthetic Judgment Ability (Bamossy et al., 1983), and the Centrality of Visual Product Aesthetics scale (CVPA, Bloch, Brunel, & Arnold, 2003). Yet, each of these measures is inadequate for the study of aesthetic sensitivity in diverse settings and for diverse research purposes. For example, the VAST consists of a series of 50 pairs of paintings. While the measure has good internal consistency (r = .70, Götz et al., 1979) and has been evaluated in various cross-cultural studies (Chan, Eysenck, & Götz, 1980; Eysenck et al., 1984; Frois & Eysenck, 1995; Iwawaki, Eysenck, & Götz, 1979), the stimulus materials are works of art, which means that the assessment of aesthetics in immediate environments is not likely to be captured by this method. Additionally, the evaluation of 50 pairs of stimuli is rather time consuming and best used in experimental settings, not field research.

Bamossy et al.'s (1983) Aesthetic Judgment Ability measure was designed within a cognitive development framework. The measure examines how aesthetic judgments are influenced by developmental stages, which includes the assumption that there are different stages of aesthetic judgment that develop over time. The development consists of learning and understanding criteria that are relevant for the judgment of aesthetic objects. The test is based on the aesthetic evaluation of three different paintings. The test has good reliability and validity (see Bamossy et al., 1983). Yet, like the VAST, it uses works of art as stimuli, and is thus not adequate to assess the perception of immediate environments. Additionally, Bamossy et al.'s measure is based on a developmental framework. The aim of the present research, in contrast, is to develop a scale that allows investigating individual differences in aesthetic sensitivity towards everyday objects and the relationship of sensitivity to other psychological constructs. Finally, the CVPA was developed in the field and for the use of consumer research and is concerned with the importance that visual aspects of products have for consumers. CVPA is understood as measuring a general trait that is independent of the visual properties of the aesthetic object. Thus, the scale does not use visual stimuli but requires evaluating eleven statements about aesthetic products (e.g., "Sometimes the way a product looks seems to reach out and grab me."). It includes three different dimensions: the personal and social value of design, the ability of a person to evaluate aesthetic objects, and the valence and intensity of responses to an aesthetic object such as positive or negative feelings towards it. Internal consistency and construct validity have been demonstrated for this scale (Bloch et al., 2003). Yet, the obvious difference between CVPA and the present approach is that the CVPA does not use any visual stimuli. Like most psychological approaches the present approach understands the construct of visual aesthetic sensitivity such that it is object related and thus should be assessed by the use of visual stimuli. Another difference is that the CVPA scale is strongly oriented towards product design and consumer behavior. However, the idea of the present scale that individuals are differentially sensitive towards how an object "looks and feels" is necessarily seen as related to the idea of consumer behavior such as reflected in the item "If a product's design really "speaks" to me, I feel that I must buy it." Despite the differences, the CVPA scale is conceptually similar to the present approach in that it concerns everyday objects and how important visual aspects of objects are to an individual. Thus, the present scale and the CVPA measure are expected to show a significant correlation in a positive direction.

d. Research Aims

The aims of the present research are to develop a scale to measure visual aesthetic sensitivity towards everyday objects (Study 1), to test the reliability of the measure (Study 2) and to provide initial evidence for its validity by examining its relationship with other related constructs (Study 3).

2) Study 1 - Exploratory Factor Analysis

The aim of Study 1 was to develop a reliable and valid scale to measure visual aesthetic sensitivity. It involved reducing the number of stimuli from an initial item pool and subjecting the remaining stimuli to an exploratory factor analysis (EFA).

a. Theoretical Considerations

i. Sample size

A wide range of recommendations regarding the sample size in exploratory factor analysis (EFA) can be found in the literature. MacCallum et al. (1999) showed that these recommendations are mostly based on the misconception that the minimum ratio of sample size and number of stimuli to achieve adequate stability and recovery of population factors is invariant across studies. Instead, they suggest that N is highly dependent on several specific aspects of a study such as the level of communality of the variables and the number of factors existing in the domain of study. The quality of factor analysis solutions improves as communalities increase and as overdetermination of factors improve (where overdeterminations means that each common factor is represented by multiple variables). High communalities lead to a relatively small impact of the sample size on the quality of the solution whereas low communalities lead to a relatively high impact on sample size. Also, as communalities increase the effect of overdetermination of factors decreases. As communalities decrease, the effect of overdetermination of factors and the sample size become more important. When many or all communalities are under .5 but there is high overdetermination of factors a sample of over 100 is required. With low communalities, a small number of factors, and three or four indicators for each factor a sample size of at least 300 is recommended. However, in cases where no knowledge exists as to the expected level of communality of the variables or the number of factors present in the study, it is recommended to obtain as large a sample size as possible (MacCallum et al., 1999). Therefore in order to determine the sample size needed for the present study, the expected level of communalities, the expected number of factors and the number of variables included in the present study need to be mentioned first.

ii. Expected level of communalities

Regarding the levels of communalities, the heterogeneity of the stimuli was an issue in the present study. Communality, defined as the variance of an observed variable accounted for by the common factors in a model (Kim & Mueller, 1978) is by definition high if the common factors of a solution capture a high amount of variance of a variable. A high amount of variance can most likely be explained by the common factors if the variables are relatively homogenous. In other words, the communalities for variables used to identify the underlying construct are likely to be higher the more homogenous the variables are. Selecting homogenous variables for a study can be relatively easy if a construct is well known. In this case former research might have shown which variables are appropriate and homogenous enough to explicate an underlying construct and to produce high communalities in a factorial solution. However, not enough research has been done on aesthetic sensitivity so far to allow for inferences on appropriate homogenous stimuli in the present study. Furthermore, homogeneity of variables also depends on the type of stimuli used for the identification of an underlying construct. Using verbal items as variables offers the possibility to formulate different items in a very similar way producing relatively high homogeneity and therefore high communalities in a factorial solution. Pictorial stimuli such as those used in the present study, however, are much more complex and therefore represent much more heterogeneous material. Using pictures of objects seems to be the more appropriate way to identify an underlying construct such as aesthetic sensitivity compared to verbal items but unfortunately it implies lowered communalities because of the heterogeneity of this material. In sum, the combination of lacking knowledge in the field of aesthetic sensitivity research and the kind of material that seems to be appropriate to measure this construct, namely pictorial stimuli, leads to the expectation of low communalities for the variables in the present study. Following the recommendations made by MacCallum et al. (1999) a rather large sample size of people seems therefore appropriate here.

iii. Expected number of factors

As mentioned above pictures of objects as stimuli are much more complex stimuli than, for example, verbal stimuli. It is therefore much more difficult to determine which aspect of a stimulus will have a certain impact on a perceiver and therefore on his or her judgment of a stimulus. The stimuli chosen in the present study were selected such that they are expected to be influenced by the aspect of “beauty”. More precisely, they were supposed to be either “ugly” or “beautiful” stimuli. Assuming that the terms “ugly” and “beautiful” represent two ends of the same dimension of a stimulus one would expect a factor analysis to result in a one-factorial solution. Alternatively one could argue that the judgment about an object being “beautiful” might be different from a judgment about an object being “ugly”. A person might pay attention to objects being more or less beautiful and therefore being influenced by the beauty of his or her environment but not paying attention to ugly objects. On the other hand one could imagine a person to pay attention to objects being more or less ugly and therefore being influenced by the ugliness of his or her environment but not having sensitivity for objects being more or less beautiful. Following this idea, one might expect the analysis to result in a two-factorial solution in which all “ugly” stimuli load on one factor and all “beautiful” stimuli load on another factor. A third possibility would be that different people might pay attention to different aspects of an object in judging its aesthetic value resulting in several common factors representing different aspects of the objects being considered when making an aesthetic judgment. In this case a multifactorial solution with only a few items loading on each factor would be expected. However, because in practice a solution with more than a certain number of factors does not seem to produce a readily interpretable and theoretically sensible pattern of results given the relatively small research background with aesthetic sensitivity on one hand and pictorial stimuli in this context on the other, the number of factors is expected not to exceed a maximum of five factors.

iv. Number of stimuli included in the study

As mentioned above research suggests that EFA procedures provide more accurate results when common factors are overdetermined and that the quality of a factor solution will improve as overdetermination of factors improves. Therefore each factor should be represented by multiple variables (Fabrigar, Wegener, MacCallum, & Strahan, 1999; MacCallum et al., 1999). Comrey and Lee (1992) recommended at least five times as many variables as factors for overdetermination. Assuming that a maximum number of interpretable common factors found in this study might not exceed five, the total number of stimuli that should be included in the study following the criterion of overdetermination would be at least 25. A practical issue was not to include too many stimuli in the study because people asked to participate in the study might not agree to participate or might withdraw from participation if judging the pictures of objects would take too much of their time. Also the participant's judgments might become unreliable (tiring effects) if they would have to judge to many stimuli.

In sum, the fact that the nature of the stimuli used in this study made it difficult to decide how many factors might emerge and that because of the nature of the stimuli communalities might be expected to be rather low, it seemed important to (a) overdetermine factors by choosing a relatively high number of stimuli without choosing an impractically high number of stimuli and (b) to decide for a rather large sample. MacCallum et al (1999) suggest a sample size of 100 participants when communalities are consistently low if there is a high overdetermiation of factors. A sample size of at least 300 participants is suggested when communalities are low and a small number of factors with just three or four indicators for each factor is expected. Finally a number of 36 stimuli were included in the study so that a number of factors up to five would be highly overdetermined. Expecting rather low communalities and a small number of factors the sample size for this study was decided to be at least $N = 300$.

b. Material, Participants, and Procedure

i. Number and type of stimuli

The initial pool of stimuli consisted of 80 stimuli intuitively chosen and reduced to 36 stimuli by the investigator of this study. Of the 36 stimuli, 18 stimuli were supposed to represent rather ugly objects and 18 stimuli were supposed to represent rather beautiful objects. The stimuli were found by surveying objects on the Internet, in catalogues of companies selling objects, and, finally, by taking pictures of objects in shops and apartments. The 36 stimuli finally included in the study were partly taken from the Internet and partly taken from a set of photographs of objects. They include objects such as a sofa, a bag, a table, a clock, a teapot and cutlery. Modifications to the pictures were made to achieve a standardized form of presentation: All stimuli should show purely the target object and not any additional objects. The pictures were thus modified using Adobe Photoshop 6.0 such that only the target objects were shown on a white background and in comparable size. The quality of the pictures was controlled, using only pictures with high quality so that quality differences could not interfere with the judgment of the shown object. For some stimuli colors were also changed to make the shown objects more “ugly” or more “beautiful.”

ii. Participants and procedure

The sample consisted of N = 308 participants, 168 females and 139 males (one person did not report his or her gender) between 16 and 77 years of age (mean age: 33.91 years; more demographic information can be found in Table 1, Appendix A). Participants were recruited in public places in downtown Montreal, such as at outdoor festivals, in downtown city parks, etc. Individuals or small groups of people (up to three individuals) were randomly approached and asked whether they would be willing to participate in a short survey. They were told that participation would simply involve looking at 36 different pictures of objects and rating how ugly or beautiful they think the objects are and that, for statistical purposes, they would be asked some demographic questions. The questionnaire material was available in English (Appendix B) and French, so that English- and French-speaking participants could participate in their native language. If individuals agreed to participate, they were given printouts of the stimuli. The printouts were available in three different random orders, from which one per participant was randomly picked. On the left side of each page stimuli were printed in a size such that four stimuli fit on each page. On the right side of each stimulus a 7-point Likert-type rating scales. Participants were asked to rate the aesthetic value of each object on such a rating scales. More precisely, they were asked for their affective reaction when seeing the pictures. Responses ranged from 0 (labeled “Urghhh, this is pretty ugly” with an icon showing a sad face) to 6 (labeled “Uihhh, this is very beautiful” with an icon showing a happy face), with the numbers 2 to 5 in between. Subsequently, participants were asked to fill out a demographic information sheet for statistical purposes. On the demographics sheet, participants were asked whether they would be willing to rate the pictures of objects again about two weeks later (in order to establish test-retest reliability) and were told that the pictures would be mailed to them together with an addressed and stamped return envelope. If they agreed, they were asked to fill in a code that allowed the researchers to connect the information from the test and retest. The name and mailing address of those who agreed were recorded on a separate sheet of paper.

c. Item Reduction and Reliability Testing

This stage of the scale development involved scale refinement and reliability testing for the 36 stimuli. In the following section, the decision processes leading to the retention or elimination of stimuli are described. Twenty stimuli were retained from the initial stimulus pool of 36.

i. *Clarity of stimuli*

Two stimuli were excluded from the data analysis because a large number of participants had difficulties identifying the objects shown in the pictures.

ii. *Analysis of stimulus distribution*

One criterion of item elimination was a skewed item distribution. The goal here was to retain only items that had a sufficiently wide distribution, or in other words would not elicit a limited range of responses. Four stimuli were eliminated because of their skewed distributions (defined as having 152 (49%) or more of the 308 participants endorse the very ugly (0) option for the particular stimulus). Another stimulus was eliminated because it showed a dichotomous distribution (defined as having 37 (8.8%) of the 308 participants endorse a value of 0, 1 or 2 and between 66 (21.4%) and 74 (24%) of the participants endorse each of the other values for this stimulus). A total of seven stimuli were eliminated, leaving 29 stimuli for further analysis.

iii. Principal axis factor analysis

The goal of the common factor analysis is to understand the structure of correlations among measured variables by estimating the pattern of relations between the common factor(s) and each of the measured variables (Floyd & Widaman, 1995). The basic idea of this study was to develop a scale assessing the latent construct of visual aesthetic sensitivity measured by complex visual stimuli. Given that no strong assumptions about the number of factors could be made, the exploratory factor analysis seemed to be the appropriate method to determine an appropriate number of factors and the factor loadings for the given set of stimuli for the data. Therefore, the remaining 29 stimuli were subjected to an exploratory factor analysis using the principal axis estimation method (PFA) for the full sample of N = 308 participants in SPSS. The analysis was set to extract all factors with eigenvalues over 1 (Cattell, 1966). Missing values were treated listwise. The Kaiser-Meyer-Olkin (KMO) test of sampling adequacy was used to determine the appropriateness of factor analysis and indicated with a level of .84 that the correlation matrix was appropriate for such an analysis (Tabachnick & Fidell, 2001). Based on the analysis of the loadings of the unrotated factors a further reduction of stimuli was conducted. Stimuli loading equally or very similar (max. difference of 0.057) on at least two factors were deleted stepwise. This method resulted in the deletion of seven stimuli. The exploratory factor analysis was then repeated with the remaining 22 Stimuli. Four factors with eigenvalues over 1 were extracted from the matrix, explaining 40.28% of the variance. The eigenvalues for the third and fourth factor, however, were only 1.22 and 1.08. An inspection of the scree plot indicated that either two or three factors should be retained. The analysis of the factor loadings showed that only two stimuli were loading on factor 3 (both with a loading of -.55). Because generally three variables per factor are needed to identify common factors (Floyd & Widaman, 1995) and because the two stimuli could not be interpreted in the sense of a common third factor here, only the first two factors (eigenvalues 4.46 and 3.54) were retained. Given that the two stimuli loading on the third factor did not load high enough (over .4) on one of the first two factors these two stimuli were also eliminated from the analysis. The factor analysis was then repeated with the remaining 20 stimuli including a varimax rotation in order to increase the interpretability of the two factors. (An oblique solution [promax rotation] returned a solution with orthogonal factors so that finally an orthogonal solution seemed to be appropriate.) All items loaded above .4 on one of the two factors (see Table 2, Appendix A for exact factor loadings) ranging from fair (.49) to excellent (.73) loadings on Factor 1 and from poor (.42) to very good (.67) loadings on Factor 2 (see Comrey & Lee, 1992 on criteria for "poor" to "very good" loadings). The secondary loadings were all acceptably low. The two factors extracted here explained only 33.51% of the total variance (see Table 3, Appendix A for initial and extracted communality estimates). The eleven stimuli of the first factor were all stimuli that had been preselected as showing “ugly” objects whereas the nine stimuli of the second factor were all stimuli that had been preselected as showing “beautiful” objects. The mean ratings of the stimuli loading on the first factor ranged between 1.25 (SD = 1.47) and 2.35 (SD = 1.91), the mean ratings of the stimuli loading on the second factor ranged between 3.02 (SD = 1.55) and 3.77 (*SD* = 1.42), illustrating that all stimuli loading on Factor 1 were on average perceived as uglier than stimuli loading on Factor 2.

iv. *Internal consistency*

As a measure of internal consistency, Cronbach's coefficient alpha (Cronbach, 1951) was calculated separately for the two factors and for the entire scale. For Factor 1 consisting of eleven stimuli, the internal consistency was $\alpha = .85$ with the highest inter-item correlation being $r = .61$ and the lowest inter-item correlation being $r = .19$. For Factor 2, consisting of nine stimuli, the internal consistency was $\alpha = .80$. The highest inter-item correlation for the second factor was $r = .42$, the lowest inter-item correlation was $r = .15$. The analysis conducted on the 20 item scale demonstrated also good internal consistency with $\alpha = .80$. The highest inter-item correlation was $r = .62$, the lowest inter-item correlation was $r = -.19$. Items loading on Factor 1 correlated negatively with some of the items loading on Factor 2, but none of these correlations exceeded a value of .19. The magnitude of Cronbach's alpha coefficient suggested that the two factors were internally consistent. The two factors were not significantly intercorrelated ($r = .03$).

v. *Test-retest reliability*

To assess the performance of the aesthetic sensitivity scale in terms of test-retest reliability, the scale was administered again two weeks after the initial assessment. The stimuli were again available as print outs in three different random orders showing the 36 pictures of different objects. This time one of the three versions was sent to the participants by mail together with an addressed and stamped return envelope. Again, the stimuli were printed on the left side of a page while a 7-point rating scale was printed beside each stimulus. Subjects were asked to rate them in the order they appeared in the packet. Half of the $N = 308$ participants of the first study agreed at time 1 to participate in the retest and provided their addresses. Responses were received from $N = 104$ individuals resulting in a response rate of 67.53%. Four responses could not be matched with any respondent at t1. Test-retest reliability was assessed separately for the two factors and for the aesthetic sensitivity scale overall. The correlations between test and retest responses were $r = .79$ for the first factor and $r = .84$ for the second factor, demonstrating high stability of the construct as assessed with the two dimension of the scales. Test-retest reliability for the aesthetic sensitivity scale as a whole was also high (r = .74). Thus it can be assumed that the scale is measuring a rather stable characteristic.

3) Study 2 - Confirmatory Factor Analysis

Confirmatory factor analyses (CFA) were conducted using the data from the validation study (N = 118) to examine whether the factorial structure found in Study 1 (two factors with 9 and 11 items, respectively) replicates. Specifically, we compared the fit of a one- and a two-factor model by using LISREL 8.30 (Jöreskog & Sörbom, 1999) software.

a. Material, Participants and Procedure

i. *Material*

The CFAs were conducted using data collected for the validation study (Study 3, see below). For this study each of the 20 pictures was printed on a 9 x 9 cm card and the cards were laminated. Each card was labeled with a number that was randomly chosen and printed on the back of the card. Answer sheets consisted of a table with two columns, the left one providing blank boxes for recording the number of the card and the right column contained the same 7-point rating scale as used in Study 1.

ii. *Participants and procedure*

The sample consisted of 97 female and 21 male German psychology students between 19 and 50 years of age (mean age: 24.6 years; SD = 5.9). The students participated in groups of 5 to 10 persons in the study for extra course credit. Only the aspects of the procedure of data collection that are important for the present study are reported here. For further description see Study 3. Each participant was provided with a pack containing the 20 laminated cards each showing a picture of one object on one side and a number on the other side, a response sheet and a pen. In each pack the cards appeared in a random order. The packs were placed in front of the participants with the upper side down so that the stimuli themselves could not be seen. Before they were asked to rate the stimuli, the experimenter explained the procedure. In order to become familiar with the stimuli participants were asked to briefly look at each stimulus included in the pack while keeping them in the given order. They were then asked to put the pack of cards back on the table with the numbers on the upper side, to fill in the number of the first stimulus into the response sheet, to turn the first stimulus card around and to make their aesthetic judgment about the stimulus on the 7-point Likert-type rating scale ranging from 0 ("Urghhh, this is pretty ugly" with an icon showing a sad face) to 6 ("Uihhh, this is very beautiful" with an icon showing a happy face). Participants were asked to evaluate their affective reaction to the stimulus. After rating the first stimulus participants were asked to put the stimulus card back on the table with the picture facing down and to repeat the described procedure with each of the stimulus cards. Once a stimulus was evaluated and put back on the table participants were not allowed to look at it again. To assess the performance of the aesthetic sensitivity scale in terms of test-retest reliability, the scale was administered again two weeks after the initial assessment.

b. Confirmatory Factor Analysis and Reliability Testing

i. Confirmatory factor analysis

Confirmatory models were examined and estimated using LISREL 8.30 (Jöreskog & Sörbom, 1999). In accordance with the results of Study 1, a two-factor CFA model with each indicator loading on one of the two underlying latent variables was specified. The model was specified as follows: Nine variables were specified to measure the latent variable "ugly", eleven variables were specified to measure the latent variable "beauty". The two constructs were assumed to be independent. To set scales for the latent variables, the loadings of the two variables that loaded most heavily on the factors of the EFA were each fixed to 1.0. Values of selected fit indices are reported in Table 4.

As indicated in Table 4, a comparison with a one-factor model showed that the overall fit of the two-factor model is significantly better (327.05 – 243.70 = 83.35, df = 1) than that of a one-factor model for these data. The relative χ^2 (χ^2/df ratio) for the two-factor model with a value of 1.43 is favorable compared with the value of 1.91 for the one-factor model.

Table 4. Goodness of fit summary

Model	χ^2	*df*	χ^2/df	RMSEA	CFI	NNFI	SRMR
two-factor CFA	243.70**	170	1.43	.061	.75	.72	.11
one-factor CFA	327.05**	171	1.91	.088	.60	.55	.10
single-factor vs. two-factor	83.35**	1					

Note: *df* is degrees of freedom, RMSEA is Root Mean Square Error of Approximation (Steiger, 1990), CFI is Comparative Fit Index (Bentler, 1990), NNFI is Non Normed Fit Index (Bentler & Bonetts, 1980), and SRMR is standardized Root Mean squared Residuals (Jöreskog & Sörbom, 1986). The critical χ^2 for 1 degree of freedom at the .01-level is 6.64.

Furthermore, indices with more standardized metrics like the Non Normed Fit Index (NNFI) indicate a preference for the two-factor model. The two-factor CFA model's Root Mean Square Error of Approximation (RMSEA) was .061, the Comparative Fit Index (CFI) was .75, the NNFI was .72 and the standardized Root Mean squared Residual (SRMR) was .11. Thus, the two-factor CFA model demonstrates satisfactory fit of the data in the relative χ^2 (χ^2/df ratio) and an adequate fit in the RMSEA.

ii. *Internal consistency*

The internal consistency was calculated for this second sample, again separately for the two factors. For Factor 1 the internal consistency was α = .68, for Factor 2 it was α = .73. The highest inter-item correlation was r = .49 for Factor 1 and r = .48 for Factor 2, the lowest inter-item correlation was r = -.10 for Factor 1 and r = .08 for Factor 2. The analysis conducted on the 20 stimulus scale lead to a value of α = .51. The highest inter-item correlation was r = .52, the lowest inter-item correlation was r = -.33. Again, the magnitude of Cronbach's alpha coefficient suggested that the two factors were internally consistent. The two factors were moderately intercorrelated with each other in this sample (r = -.316, p = .000).

iii. *Test-retest reliability*

From the initial sample of 118 participants, 117 rated the stimuli a second time two weeks later following the same procedure as described above. Again, test-retest reliability was assessed separately for the two factors and for the aesthetic sensitivity scale overall. The correlations between test and retest responses were r = .78 for the first factor and r = .76 for the second factor, again demonstrating high stability of the construct as assessed with the two dimensions of the scale. Test-retest reliability for the aesthetic sensitivity scale as a whole was acceptable (r = .65). Thus, this sample confirms that the scale is measuring some rather stable quality.

In sum, the results of the confirmatory factor analysis showed that the two-factor-model is superior to a single-factor model. Fit indices for the relative χ^2 (χ^2/df ratio) were satisfactory and the RMSEA had adequate fit. The CFA thus confirmed the solution found in the EFA. Additionally, internal consistency and retest-reliability for the two factors and the scale overall could also be demonstrated in this sample.

4) Study 3 - Convergent Validity and Relationships with Other Measures

The objective of Study 3 was to provide initial evidence for convergence and divergent validity of the measure. One goal was to establish convergent validity with the present measure with respect to other measures of aesthetic perception, individual differences in using visually oriented information and self-reports on aesthetic sensitivity. Another goal was to investigate the relationship between the ugly and the beauty scale with characteristics of people's living environment. This was measured by assessing the frequency of visiting art museums (see Child, 1965) and attributes of their living space (see Bourdieu, 1979). The study also served to examine the scale's sensitivity toward socially desirable responding.

a. Participants, Procedure, and Measures

i. *Participants and procedure*

A description of the sample and the first part of the procedure of this study can be found in the description of Study 2. Altogether, N = 118 German psychology students participated. (For a detailed description see Study 2). Once finished with rating the stimuli, participants were provided with a survey containing measures of nine related constructs followed by some demographic questions.

ii. *Scoring for ugly and beauty scale*

Ratings on the 7-point rating scales were summed up and divided by the numbers of stimuli to build the ugly and beauty scale, respectively. The possible range for the scores on both scales is thus 0 to 6. Prior to summing up the items, the scores of the ugly scale were reversed so that a high total score indicates high aesthetic sensitivity. Although the terms *ugly* and *beauty* scale might suggest that these two aspects are opposites of one dimension (and thus should be strongly correlated), they in fact emerged as distinct dimensions. Consequently, a single scale value was not computed that included the ratings on both scales. Hypotheses will be tested for the ugly and the beauty scale separately. In addition to the ugly and beauty scale the following measures were used:

iii. *Test of Aesthetic Judgment Ability (Bamossy et al., 1983)*

It was assumed that a person who is more aware of relevant features of aesthetic objects in terms of art work is also more aware of relevant features of everyday objects. Thus, a significant correlation in a positive direction between the Test of Aesthetic Judgment Ability and the present scales was expected. However, because the Test of Aesthetic Judgment Ability refers to works of art, and therefore a priori more to beautiful objects, a stronger correlation was expected for the beauty scale than for the ugly scale.

iv. *Scale for Centrality of Visual Product Aesthetics (Bloch et al., 2003)*

Individuals who receive high scores on the CVPA measure are those who are sensitive to the visual aesthetics of objects and consequently should also score high on the present scales. Similarly to the Test of Aesthetic Judgment Ability, the CVPA scale addresses mainly beautiful objects. Consequently, again a higher positive correlation was expected between the CVPA and the beauty scale.

v. *The visual dimension of the Style of Processing Scale (Childers, Houston, & Heckler, 1985)*

Like the Test of Aesthetic Judgment Ability and the CVPA measure the SOP scale was developed in the field of consumer research. The SOP focuses on individual differences in the preference to engage in visual versus verbal processing of information. The basic idea of the construct is that individuals differ in their preference for using visually versus verbally oriented information across various situations. A preference for visual or verbal processing leads an individual to prefer one kind of stimulus over the other if both kinds of stimuli are available. If only one kind of stimulus is available, an individual tends to represent information in his or her preferred modality. The SOP scale is a 22-item scale and uses 5-point Likert scales. The final score of a person ranges from visual to verbal processing, with low scores indicating a preference for visual processing. As mentioned by the authors (Childers et al., 1985) it is possible to only assess either the verbal or the visual dimension of the scale. Because the present research uses only visual stimuli, only the visual dimension of the SOP scale was used. Internal consistency and construct validity have been demonstrated for this scale (Childers et al., 1985; Heckler, Childers, & Houston, 1993). The internal consistency of the visual component was found to be $\alpha = .86$ (Childers et al., 1985). In line with earlier findings (Bloch et al., 2003; Brunel, 1998) it is suggested that individuals who show a stronger preference for visual processing are more likely to score high on the aesthetic sensitivity scales. When seeing objects, a person with a stronger tendency to visually process information might form better representations of the objects in his or her memory and thus form more discriminative judgments when evaluating aesthetic differences between objects.

The three measures just described only were available in English. Consequently, they had to be translated for use in the present study (see Appendix C for the German translations). For the internal consistency of the translated scales in the present sample see Table 5.

vi. *Self-report measures*

In addition to the above measures, three single-item-measures were developed to evaluate convergent validity. They consisted of three statements, which participants were asked to rate on a 7-point Likert-type rating scale ranging from -3 to +3. The statements were (a) “I rate my ability to judge the aesthetic values of objects as…,” b) “I sometimes enter a room which I find so ugly that I want to leave it immediately," and c) "I can rarely tell with certainty if I find something ugly or beautiful." The labels for the rating scale ranged from "very bad" to "very good" for the first question, and for the second and third question from "highly disagree" to "highly agree." It was assumed that individuals with higher aesthetic sensitivity would rate themselves as significantly more aesthetically sensitive, that they would more strongly agree with the statement that they sometimes enter a room that they find so ugly that they want to leave it immediately and that they would report feeling more certain when judging the aesthetics of objects.

vii. Visits to art museums

Four items assessed a person's exposure to art museums. It was assumed that individuals high in aesthetic sensitivity would visit art museums more frequently than individuals low in aesthetic sensitivity. The four items were the most popular art museums in Berlin. Participants were asked to indicate for each museum whether they know it and how often they had been there in the last year. Responses were given on a 4-point Likert-type rating scale with the response options "don't know it" (0), "know it, but have not visited it yet" (1), "visited once" (2), "visited several times" (3).

viii. Living space

Eight items assessed aesthetic attributes of a person's current and ideal living space and were summarized into an index. These are the items warm (warm), comfortable (komfortabel), convenient (praktisch), neat (gepflegt), conventional (konventionell), functional (sachlich), dark (düster) and bright (hell). These attributes were categorized as functional (convenient, neat), stylistic (conventional, functional, warm, comfortable) or atmospheric (dark, bright) aspects. Participants were asked to (a) indicate which of these attributes describe their current living space and (b) choose the five attributes that describe best how they ideally would like to design their living space. A positive relation with aesthetic sensitivity was expected for the attribute "bright". For all remaining attributes a significant negative correlation was expected.

ix. Social desirability

To assure that the present scale is not susceptible to social desirability, correlations between responding to the ugly and beauty scale with social desirability were assessed. Social desirability was measured using the German version of the Social Desirability Scale-17 (SDS-17, Stöber, 1999, 2001). Building upon the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960), the SDS-17 measures people's need to present themselves in a favorable light. The measure is widely used to assess the tendency to endorse the items of a self-report measure in a socially desirable way. The scale consists of 16 items (one item was deleted from the original scale). Each item has to be evaluated as "right" or "wrong". Internal consistency (α = .75) and construct validity has been demonstrated for this scale (Stöber, 1999, 2001).

b. Results

Table 5 (Appendix A) presents the means and standard deviations, the correlations of the above measures with the ugly and the beauty scale, and Cronbach's coefficient alpha as indicator of internal consistency for the different scales in the present sample. All measures showed satisfactory to good internal consistency ranging from α = .53 for the Test of Aesthetic Judgment Ability to α = .87 for the CVPA measure.

i. Convergent validity

One goal was to examine how the ugly and the beauty scale are related to the Test of Aesthetic Judgment Ability. As can be seen in Table 5 (Appendix A), a significant positive correlation between scores on the beauty scale and scores on the Test of Aesthetic Judgment Ability was found (r = .19, p = .04), suggesting that the two scales overlap in their assessment of aesthetic sensitivity. As anticipated, the correlation found between scores on the ugly scale and scores on the Test of Aesthetic Judgment Ability (r = .15, p = .11) was not significant.

Looking at the relationship between the ugly and the beauty scale and the CVPA measure, Table 5 (Appendix A) shows that there is no significant correlation between the measures. However, an examination of the correlations with the three different dimensions of the CVPA measure showed significant correlations between the dimension 'value' and the ugly ($r = .18$, $p = .05$) as well as the beauty scale ($r = .18$, $p = .05$). This suggests that there is some overlap in the aesthetic sensitivity as assessed by the value dimension of the CVPA and the two present scales.

For the visual dimension of the SOP scale, the scores were reversed so that high scores indicate a high preference for visual processing. A significant correlation with the beauty scale but not with the ugly scale emerged (see Table 5, Appendix A). Participants with high scores on the visual dimension of the SOP scale also score higher on aesthetic sensitivity as measured with the beauty scale ($r = .19$, p = .04). The correlation for the ugly scale was not significant ($r = .11$, $r = .25$).

Next, the scales' associations with the self-report measures were examined. The statement about judgment certainty was reverse-coded for the analysis. Table 5 (Appendix A) shows that the scores of the beauty scale are significantly correlated with the self-rating of aesthetic sensitivity ($r = .22$, $p = .02$). Providing further support for these findings the same relationship between the self-rating of aesthetic sensitivity was found for the North American sample ($r = .13$, $p = .02$). In the German sample scores of the beauty scale are furthermore positively correlated with endorsements of the statement that a person sometimes enters a room which he or she finds so ugly that he or she wants to leave it immediately ($r = .24$, $p = .01$) and with judgment certainty ($r = .19$, $p = .04$). For the ugly scale, scores are only significantly correlated with the self-rating in aesthetic sensitivity ($r = .21$, $p = .03$). In the North American sample participants were also asked how important it is for them that things around them look nice. This variable was significantly correlated with the beauty scale ($r = .15$, $p = .01$) and close to significance for the ugly scale ($r = .11$, $p = .06$).

Together, these results support the expected pattern of associations, particularly for the beauty scale. For this scale, convergent validity is found with the Test of Aesthetic Judgment Ability, the value dimension of the CVPA measure, the visual dimension of the SOP scale and all self-reports on aesthetic sensitivity. For the ugly scale convergent validity was found for the dimension of the CVPA measure and self-reported aesthetic sensitivity. The correlations between the ugly scale and the other measures were not significant.

ii. *Relation to other measures*

Scores for visiting art museums were calculated by summing up across the four items, resulting in a range of possible scores from 0 to 12. Table 5 (Appendix A) shows a significant correlation with the beauty scale (r = .24, p = .01). As expected, individuals who score high on the beauty scale also visit art museums more frequently. Again the correlation with the ugly scale was not significant (r = .12, p = .19).

The index for both aspects of the living space, the current and the ideal living space, was built by counting which attributes a person had chosen to describe his or her current and ideal living space. The item “bright” was reversed for building the index because it was supposed to be positively correlated with aesthetic sensitivity. For all other items negative correlations were expected. Table 5 (Appendix A) shows negative correlations of the index with the beauty scale (r = -.22, p = .02) and the ugly scale (r = -.29, p = .001). For the ideal living space a significant negative correlation was found with the beauty scale (r = -.27, p = .00). Individuals with high scores on the beauty scale consider functional aspects and certain styles of their living space as less important and prefer a bright atmosphere. The correlation with the ugly scale was not significant (r = -.16, p = .09).

iii. *Social desirability and socio-demographic characteristics*

The scales’ sensitivity to social desirable responding was examined (see Table 5, Appendix A). Scores on the ugly and the beauty scale were not significantly related to the tendency to respond in a socially desirable way (r = -.04, p = .67*;* r = -.01, p = .95). Scores also did not correlated with socio-demographic characteristics such as age, gender and income (all p > .30), except for age, which correlated with the ugly scale (r = .16, p = .08). Controlling of age in the correlations reported above did, however, not notably change the results.

5) Discussion

This chapter described the development and psychometric properties of a scale for measuring visual aesthetic sensitivity towards everyday objects. Results of the exploratory factor analysis (Study 1) showed that the scale consists of two rather independent factors, one labeled “ugly” and the other “beauty.” Results of the confirmatory factor analysis (Study 2) confirmed this two-factor solution. These findings suggest that affective responses do not range on one dimension from beautiful to ugly but rather consist of two orthogonal dimensions representing affective responses to beautiful objects and ugly objects, respectively.

Results further showed that the scales are reliable and give initial evidence for convergence and divergent validity of the measure. Scores on the beauty scale showed convergent validity with all measures except for the CVPA measure, for which only the value dimension showed convergent validity. Furthermore, visits to art museums and the attributes of living spaces were found to be significantly correlated with the beauty scale. For the ugly scale convergent validity was found for the value dimension of the CVPA measure and for self-reported aesthetic sensitivity. In addition, attributes of the current living space were significantly correlated with scores on the ugly scale. Both aesthetic sensitivity scales were not correlated with the tendency to respond in a socially desirable way. Altogether, these results show that the beauty scale is a reliable and valid instrument to measure aesthetic sensitivity. For the ugly scale further research is needed before the measure can be used to assess aesthetic sensitivity.

a. Limitations

Some limitations of the present studies should be noted. The first limitation refers to the comparably small amount of variance that is explained by the exploratory factor analysis. A commonly used criterion in EFA holds that the explained variance should account for at least 50% of the total variance in the data (e.g., Steiner, 1994). The amount of variance explained by the two factors in Study 1 (33.5%) consequently seems rather small. One explanation of the comparatively small amount of explained variance is the heterogeneity of the visual stimuli. By nature, visual stimuli are much more complex and therefore differ much more from each other than verbal items (e.g., Bloch et al., 2003). Each stimulus varies in a variety of aspects such as form, color, texture, and so on. The communalities of each stimulus and consequently the amount of variance explained by the factors tend to be rather low for scales with heterogeneous stimuli such as pictures of objects. An exception would be stimuli that are constructed to be homogenous such as polygons. Approaches that are similar to the present one using complex visual stimuli also have found rather small amounts of explained variance. For example, Eysenck (1940) analyzed eighteen sets of pictures each containing on average fifteen pictures. The stimuli included reproductions of portraits, drawings from Claude Lorrain, Japanese paintings, photographs of vases, masks, flowers, pottery, plates and other objects. The two factors he found also accounted together only for 34.3% of the variance. Thus, the amount of variance explained in the present study seems equivalent to other scales using visual stimuli.

Another limitation of the present results is related to the goodness of fit indices found in the confirmatory factor analysis. One reason for the lack of a better fit might be the relative small sample size in combination with the relative high number of variables used in the analysis. Because the data used for the CFA were not primarily collected for this purpose, sample size was not determined by criteria of conducting CFAs. Future research should examine whether the results of the CFA can be replicated in a larger sample. A second reason for the relatively low goodness of fit is the comparably small amount of explained variance in the EFA found in Study 1. EFA and CFA have certain differences in their analytic techniques. The aim of EFA is to identify factors that account for significant amounts of variance in the data and this is what EFA mainly focuses on in its goodness of fit measures. CFA takes the found factors into account but its goodness of fit assessment is based on the remaining variance (Floyd & Widaman, 1995). The relatively low goodness of fit measures in CFA show that significant variance remains in the data additional to the one explained by the two factors identified by the EFA. As described above, heterogeneity in the visual stimuli might have led to the comparably small amount of explained variance. Consequently a large amount of variance remains unexplained. The stimulus material, namely the complexity of the visual stimuli, might thus be responsible for the lack of a better fit in the CFA. Nevertheless, in spite of the comparably low goodness of fit indices, Study 2 confirmed the factorial solution found in Study 1.

In sum, the results of the studies described in this chapter suggest that the heterogeneity of the stimuli used in empirical aesthetics research such as the present scale development result in relatively low goodness of fit measures in EFA and CFA compared to studies using verbal or very simple visual stimuli. These results illustrate that more research is needed on criteria that are used to evaluate the aesthetic values of objects. Such criteria could be used to chose or even create stimuli that are homogenous visual stimuli but differ in their aesthetic value.

b. Sensitivity for Beauty versus Ugliness

It is argued that the two independent factors "ugly" and "beauty" reflect that aesthetic sensitivity for ugly objects is different from sensitivity for beautiful objects and that a person who is sensitive to the aesthetic value of ugly objects is not necessarily sensitive to the aesthetic value of beautiful objects and vice versa. Imagine for example a person entering a room full of beautiful objects. This person might be sensitive to the degree of beauty of the exposed objects. The same person, however, entering a room full of ugly objects, might not be sensitive to aesthetic differences between these objects at all. Therefore, a person who is interested in and focusing only on beauty in his or her environment and does not pay attention to ugliness might be sensitive for beautiful objects but not for ugly ones. The results of the research presented in this chapter suggest that beauty and ugliness are not two ends of the same dimension but two independent dimensions.

c. Why higher Convergent Validity for the Beauty Scale?

Study 3 found that scores on the beauty scale showed convergent validity with all measures except for the CVPA measure, for which only the value dimension showed convergent validity. For the ugly scale convergent validity was only found for the value dimension of the CVPA measure and for self-reported aesthetic sensitivity. One reason for the higher correlations found for the beauty scale could be that the other instruments also focus on the perception of aesthetically pleasing objects. Given that the ugly scale appears to measure an independent dimension of aesthetic sensitivity rather than the opposite of the aesthetic sensitivity for the beauty, the selected scales were not sensitive measures of convergent validity for the ugly scale.

The beauty scale showed a significant, but only small correlation with the Test of Aesthetic Judgment Ability (Bamossy et al., 1983). As mentioned earlier, although both scales measure aesthetic sensitivity, several differences between the measures exist. First, this measure uses works of art and is thus not suitable to assess perceptions of the immediate environment. Secondly, it is based on a developmental framework. The aim of the present research, in contrast, was to develop a scale that allows one to investigate individual differences in aesthetic sensitivity towards everyday objects and the relationship of different levels of sensitivity to other psychological constructs. Third, the Test of Aesthetic Judgment

Ability requires agreeing to statements about art works whereas the present scale asks for the affect that looking at everyday objects evokes. However, even though the two scales differ in the underlying theoretical frameworks, stimuli and aesthetic responses, they still show a significant correlation, suggesting that they are measuring at least in part the same construct, namely people's visual sensitivity towards aesthetic objects.

An explanation for the lack of a significant correlation with the overall CVPA score can be found in the items of the measure itself. Specifically, the items are strongly oriented towards product design and consumer behavior and are supposed to measure aesthetic sensitivity independent of objects. The scale uses verbal items that include purchasing behavior and possession of products. In contrast, the construct of aesthetic sensitivity as measured by the present scale only refers to sensitivity in perceiving differences in the aesthetic value of objects, but does not require respondents to make any implications about purchasing or possessing them. Thus, the two scales appear to measure quite different aspects of the visual aesthetic sensitivity construct. The lack of a significant correlation might reflect the different orientations of the measures. However, the overlap between the value dimension of the CVPA scale and the beauty scale might reflect a form of pleasure that people derive from dealing with beautiful objects. Individuals high on the value dimension believe that dealing with beautiful objects influence their daily lives positively. These individuals also show higher visual aesthetic sensitivity towards beautiful everyday objects. Consequently, one might assume that individuals who are more sensitive towards beautiful objects also perceive more pleasure in and find it more important to deal with beautiful objects in their everyday life. This interpretation is also supported by the correlations found for the self-report measures, the visits to art museums and the current and ideal living space with the beauty scale. Individuals who score high on the beauty scale perceive themselves as more aesthetically sensitive, they more often experience situations in which they want to leave a room because they find it too ugly and feel more certain in making aesthetic judgments. Furthermore, aesthetically sensitive individuals visit art museums more often and they describe their current and ideal living space as less functional or oriented towards certain styles such as conventional or comfortable. Thus, overall it seems that aesthetically sensitive individuals try to surround themselves with more aesthetically pleasing objects.

Only the current but not the ideal living situation was correlated significantly with the scores in aesthetic sensitivity on the ugly scale. Thus, in sum, the beauty scale showed stronger relations to the other measures of aesthetics used in the present study than the ugly scale. One conclusion that can be drawn from this result is that beautiful stimuli are more capable of assessing aesthetic sensitivity than ugly stimuli. Considering that “aesthetic” is characterized by an appreciation of beauty (and not of ugliness), this result is not surprising.

The previous chapter described the development and psychometric properties of a scale for measuring visual aesthetic sensitivity towards everyday objects. The outcome is a reliable and somewhat valid measure of visual aesthetic sensitivity for everyday.

A problem of the research described in the previous chapter was the choice of stimuli for scale development. The stimuli were chosen from an available pool of real-life stimuli. First, it was difficult to identify stimuli that differ in their aesthetic value because there is an *a priori* lack of relevant decision criteria. Thus, stimuli were chosen that were, to the author of this thesis, obvious exemplars of ugly and beautiful objects. This is arguably not the most rigorous procedure for choosing stimuli. Second, because of the extreme heterogeneity of the stimuli, only a comparably small amount of variance of 33.5% was explained by the exploratory factor analysis. Also the confirmatory factor analysis resulted in relatively low goodness of fit measures compared to studies using verbal or very simple visual stimuli. These results indicate that more homogenous stimuli are needed for building a reliable and valid scale of aesthetic sensitivity. The fact that a different than the classical approach to scale development used in the previous chapter is needed for constructing a reliable and valid scale for measuring visual aesthetic sensitivity is underlined by the effect sizes of the correlations found with other scales. Even though most of the correlations were significant at least for the beauty scale, they were overall very low in terms of effect size ranging from r = .04 to r =.29.

When thinking about an alternative approach to scale development in visual aesthetics the most striking question seems to be: How can homogenous stimuli be selected that still differ sufficiently in their aesthetic values when no relevant judgment criteria for the respective stimuli are available? I propose that the answer to this question is the identification of relevant criteria for judging the aesthetic sensitivity of stimuli. If such criteria can be identified, they can be used to choose stimuli that are less heterogeneous than those used in the previous chapter but still differ in their aesthetic values. Chapter 2 describes a way to identify such criteria. More precisely, the aims of the studies described in Chapter 2 are to identify relevant aesthetic judgment criteria and to use them for stimulus construction. Expert interviews are conducted to obtain information concerning the dimensions that experts use when judging the aesthetic value of stimuli. Multidimensional unfolding (MDU) is then used on non-expert data in order to identify relevant aesthetic dimensions inherent in stimuli. Finally, using the information derived from the expert interviews and the MDU studies, stimuli are constructed that vary on no other dimension than the relevant aesthetic dimensions. Such stimuli can be utilized in various research contexts dealing with empirical aesthetics. They could, for example, be used to investigate the importance of each aesthetic dimension for the overall aesthetic judgment of the given stimuli and thus allow for a better understanding of the aesthetic judgment itself. In terms of scale development the stimulus pool consists of very homogeneous stimuli which nonetheless vary considerably and systematically in their aesthetic value. Thus, the stimuli are ideal for constructing a reliable and valid measure of visual aesthetic sensitivity.