Chapter III The Use of systematically varied Stimuli for Scale Development

1) Introduction

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). The aim of the present research was to develop a scale that has good psychometric properties, is developed for diverse research contexts, is easy to administer and includes art works such as painting as well as everyday objects.

Even though aesthetic principles have been discovered that might be important for the aesthetic appeal of objects such as symmetry, balance, clarity, color, novelty and many others (e.g., Berlyne, 1963, 1970, 1974a; Boring, 1942; Koffka, 1935; Metzger, 1953), it is not yet known for everyday objects which characteristics and thus which stimuli educe a positive response in an aesthetically sensitive perceiver. Consequently, when considering everyday objects it seems rather difficult to identify stimuli that differ in their aesthetic value. To assure that the stimuli used for the present scale development represent everyday objects differing in their aesthetic value, new stimuli were constructed in prior research presented in Chapter 2. For this purpose first relevant aesthetic dimensions were identified using multidimensional unfolding (Study 5, Chapter 2). Then new stimuli were constructed, varying along the aesthetic dimensions that were identified as commonly used judgment criteria for objects of the given stimulus classes in the unfolding study (Study 6, Chapter 2). The following studies were designed to build a final scale measuring visual aesthetic sensitivity using stimuli that were constructed varying systematically on relevant aesthetic dimensions. The process of scale development included stimulus reduction, reliability (Study 1) and validity (Study 2) testing.

2) Theoretical Considerations

a. The External Standard

When aesthetic sensitivity is investigated in aesthetics research it is usually evaluated according to how much an individual's aesthetic judgment deviates from an external standard (e.g., Child, 1962; Child, 1964, 1965; Eysenck, 1988). This so-called "objective" aesthetic value of stimuli has mainly been established either by: (a) experts (e.g., Berlyne, 1971; Child, 1962), (b) calculation of averages of aesthetic judgment across participants for the stimulus, or (c) a combination of both (see Eysenck, 1988). There have been several studies showing that the group-mean approach for non-experts (viz. participants) results in similar results as the expert approach (see Eysenck, 1947) and that the degree of agreement tends to be constant over different kinds of material (Child, 1962; Eysenck, 1947). Building on this research, the external standard used in the present research is the average aesthetic judgment across participants. Consequently, the extent to which a person agrees with the average aesthetic judgment for a given stimulus is seen as the amount of his or her sensitivity. That is, the more concordant a person evaluates the stimuli with the stimuli's average judgments, the higher his or her aesthetic sensitivity.

b. Aesthetic Judgment

Aesthetic judgments and aesthetic preferences are concepts that are related to the evaluation of aesthetic objects. The evaluation of an aesthetic object is considered an aesthetic judgment when the *aesthetic value* of the object is evaluated. In contrast, an expression of a person's *relative liking or disliking* of the object is considered an aesthetic preference (Child, 1964). The present study assesses aesthetic judgments of objects, not aesthetic preferences. More precisely, participants were explicitly asked not to state their personal preference; rather they were asked how beautiful the object is in an "objective" sense (for exact wording see Appendix E).

The following studies were designed to develop a scale measuring visual aesthetic sensitivity using stimuli that were constructed in and taken from studies described previously (see Chapter 2). These stimuli vary systematically on relevant aesthetic dimensions. The process of scale development included stimulus reduction, reliability (Study 7) and validity (Study 8) testing.

3) Stimulus Reduction

The stimuli constructed in my prior studies reported earlier (Study 6, Chapter 2) were used for the present scale development. In order to construct a scale that is easy to administer, the number of stimuli had to be reduced to the minimum number necessary for effectively assessing aesthetic sensitivity.

The initial stimulus pool consisted of 34 pictures of objects which represented each of the four different object classes (i.e., paintings, offices, car interiors, and cutlery) (see Appendix D). The data from the multidimensional unfolding study (Study 6, Chapter 2) were used to accomplish the stimulus reduction. The stimuli were reduced performing pairwise comparisons for each pair of stimuli within each object class using the Wilcoxon test. The test was adjusted for Type I error in multiple testing using Holm's method (Aickin & Gensler, 1996; Holm, 1979). The decision processes leading to the retention or elimination of stimuli was based on the following criteria. A first criterion was to select the pair with the highest separation performance. A second criterion was that each dimension used to construct the stimuli should be represented by at least one stimulus pair. More specifically, pairs were chosen such that between the two stimuli of any pair (e.g., "high, high, low" – "high, low, low") one dimension changed while the other dimensions were held constant. A third criterion was that the smallest number of stimuli possible should remain for the final scale. Following these criteria, four painting stimuli, three office stimuli, four car interior stimuli and four cutlery stimuli remained in the final scale. Table 8 (Appendix A) shows the retained stimulus pairs with the respective dimensions, their levels, their separation performance and the corresponding corrected *p*-values for all four object classes.

For the office stimuli, the pair with the second highest separation performance was retained because this allowed the choice of fewer stimuli for the final scale. Additionally, choosing this pair allowed the inclusion of the original stimulus in the final scale. For the car interiors, no pair with a significant separation performance for the technology dimension was found. Thus, for this dimension a stimulus pair was retained for which one stimulus was already part of another pair and the second stimulus showed the highest possible separation performance with one of the other stimuli that were already selected for the final scale.

In sum, combining the different criteria mentioned above, the initial stimulus pool of 34 stimuli was reduced to 15 stimuli.

4) Study 7 - Exploratory Factor Analysis

The aim of Study 7 was to investigate the factorial structure and the reliability of the present scale. Thus, the remaining stimuli were subjected to an exploratory factor analysis (EFA).

a. Participants and Procedure

The sample consisted of $N = 250$ participants, 144 females and 104 males (two persons did not report his or her gender) between 15 and 99 years of age (mean age: 31.27 years; more demographic information can be found in Table 9, 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 15 different pictures of objects and rating how beautiful they think the objects are and that, for statistical purposes, they would be asked some demographic questions. The questionnaire explicitly asked them to "try not to state your personal preference, but rather how beautiful the object is in an "objective" sense". The questionnaire material was available in English (see Appendix E) 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. Participants were asked to rate the aesthetic value of each object on 7-point Likert-type rating scales. Responses ranged from 0 (labelled "not beautiful at all") to 6 (labelled "very beautiful"), 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.

b. Principal Axis Factor Analysis

The basic idea of this study was to develop a scale assessing the latent construct of visual aesthetic sensitivity measured by complex and systematically varied visual stimuli. Given that no strong assumptions about the number of factors can be made, the exploratory factor analysis seemed to be the prudent method to determine an appropriate number of factors and the factor loadings for the given set of stimuli for the data. Therefore, the 15 stimuli of the final scale were subjected to an exploratory factor analysis using principal axis estimation method (PFA) for the full sample of $N = 250$ participants in SPSS. The analysis was set to extract all factors with eigenvalues over 1 (Cattell, 1966). Missing values were deleted 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 .78 that the correlation matrix was appropriate for such an analysis (Tabachnick & Fidell, 2001). Four factors with eigenvalues over 1 were extracted from the matrix, explaining 60.20% of the variance. The eigenvalue for the fourth factor was only 1.18. An inspection of the scree plot indicated that either three or four factors should be retained. Yet, the analysis of the factor loadings showed that no stimulus loaded higher on factor four than on another factor. Thus, only three factors (eigenvalues 4.13, 2.94 and 3.21) were retained. The factor analysis was then repeated with all 15 stimuli but extracting a three-factor solution. An oblique rotation (promax) was performed on the three factors to increase their interpretability. The factor correlation matrix of the factor solutions showed that the first and the third factors were correlated ($r = .23$), suggesting an overlap in variance between the factors. Oblique rotation provides a better simple structure and more stable factor solutions in such cases and is therefore used as the basis for factor interpretation (Fabrigar et al., 1999). One item loaded .40 but all other items loaded above .52 on one of the three factors (see Table 10, Appendix A for the pattern matrix). In fact, for Factor 1 loadings ranged from fair (.40) to excellent (.83) (see Comrey & Lee, 1992 on criteria for "poor" to "very good" loadings). Loadings on Factor 2 were excellent, ranging from (.81) to (.87). Finally, Factor 3 loadings ranged from good (.58) to excellent (.80). The secondary loadings were all acceptably low. The three factors extracted

here explained 54.27% of the total variance (see Table 11, Appendix A for initial and extracted communality estimates).

The seven stimuli of the first factor (Factor 1) were the three office stimuli and the four car interior stimuli. The four stimuli of the second factor (Factor 2) were the painting stimuli. The four stimuli of the third factor (Factor 3) were the cutlery stimuli. The stimuli loading on Factor 1 – namely offices and car interiors – are stimuli that represent a room or space in which a person can move around. In contrast, the other two factors represent objects (i.e., cutlery or paintings). Thus, Factor 1 was labeled "space" and Factors 2 and 3 simply kept the names of the objects they represented, painting and cutlery, respectively.

c. Internal Consistency

As a measure of internal consistency, Cronbach's coefficient alpha (Cronbach, 1951) was calculated. The internal consistency for the space subscale was $\alpha = .87$, with the highest inter-item correlation being $r = .83$ and the lowest inter-item correlation being $r = .25$. For the painting subscale the internal consistency was $\alpha = .91$, with the highest inter-item correlation being $r = .79$ and the lowest being $r = .64$. And, finally, for the cutlery subscale the internal consistency was $\alpha = .84$, with the highest inter-item correlation being $r = .82$ and the lowest being $r = .40$. The internal consistency for the overall scale (i.e., across the three factors) was $\alpha = .81$, with the highest inter-item correlation being $r = .76$ and the lowest inter-item correlation being $r = -.06$. The magnitudes of Cronbach's alpha coefficient for the subscales and the scale overall suggested that the scales are highly internally consistent.

Table 12 (Appendix A) shows that the subscales are not significantly intercorrelated. The correlations between each subscale and the overall scale are all significant (all ps $< .01$).

d. 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 printouts in three different random orders showing the 15 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. From the $N = 250$ participants

of the first study, N = 114 agreed at time 1 to participate in the retest and provided their addresses. Responses were received from N = 65 individuals, resulting in a response rate of 57.02% . The correlations between test and retest responses were r = .69 for the space subscale, r =.65 for the painting subscale, and r = .62 for the cutlery subscale. For the overall scale the correlation between test and retest responses was r = .68. These correlation values demonstrate stability of the construct as assessed with the present scale.

5) Study 8 - Convergent Validity and Relationships with Other Measures

The objective of Study 8 was to provide initial evidence for convergent and divergent validity of the measure. Additionally, test-retest reliability was evaluated again in this study. In more detail, one goal was to establish convergent validity of the presented 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 present 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 peoples their living space (see Bourdieu, 1979). Finally, the study served to examine the scale's sensitivity toward socially desirable responding.

a. Material, Participants, Procedure, and Measures

i. Material

For this study each of the 15 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. Answering 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 containing the same 7-point rating scale as used in Study 7.

ii. Participants and procedure

The sample consisted of N = 118 participants, 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. Each participant was provided with a pack containing the 15 laminated cards each showing a picture of one stimulus 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 rating scale ranging from 0 ("not beautiful at all") to 6 ("very beautiful"). 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. Reliability Testing

i. Internal consistency

The internal consistency calculated for this second sample was $\alpha = .78$ for the space subscale, with the highest inter-item correlation being $r = .60$ and the lowest inter-item correlation being $r = .08$. For the painting subscale the internal consistency was $\alpha = .71$, with the highest inter-item correlation being $r = .60$ and the lowest inter-item correlation being $r = .13$. The internal consistency for the cutlery stimuli was $\alpha = .63$, with the highest inter-item correlation being $r = .61$ and the lowest being $r = .08$. For the overall scale the internal consistency was $\alpha = .75$. The highest inter-item correlation for the overall scale was $r = .61$ and the lowest inter-item correlation was $r = -.15$. Again, the magnitude of Cronbach's alpha coefficient suggested that the scale is internally consistent.

As Table 13 (Appendix A) shows, the space subscale correlated significantly with the cutlery subscale (r =.34, p = .00). The correlations between each subscales and the overall scale are all significant, ranging from r = .52 to r = .81 (all ps < .01).

ii. 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. The correlation between test and retest responses was r = .66 for the space subscale, r = .53 for the painting subscale and r = .67 for the cutlery subscale. For the overall scale the correlation between the test and retest responses was r = .65, again demonstrating stability of the construct as assessed with the present scale. Thus, the data provide further support that the scale is measuring a rather stable quality.

c. Validity Testing

i. Scoring

In the present study, the average aesthetic judgment across participants was used as an external standard. Thus, a person's score on the present scale was calculated as the deviation of a person's judgment from the average aesthetic judgment for each stimulus object across all factors. In other words, the deviation of a person's rating from the average judgment for each stimulus was calculated. These deviations were then summed up and divided by the total number of stimuli.

In the following the instruments for assessing convergent and divergent validity are described.

ii. Test of Aesthetic Judgment Ability

Bamossy et al.'s (1983) 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 test is based on the aesthetic evaluation of three different paintings. It has good reliability and validity (see Bamossy et al., 1983). The aim of the present research 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. 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 between the Test of Aesthetic Judgment Ability and the present scales was expected.

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

The CVPA 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 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). Individuals who receive high scores on the CVPA are those who are sensitive to the visual aesthetics of objects and consequently should also score high on the present scales.

iv. The visual dimension of the Style of Processing Scale (Childers et al., 1985)

The SOP scale 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. The final score of a person ranges from visual to verbal processing, with low scores indicating a preference for visual processing. It is possible to only assess either the verbal or the visual dimension of the scale (Childers et al., 1985). 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 et al., 1993). The internal consistency of the visual component was found to be α = .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 visual aesthetic sensitivity scales.

v. 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 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 visual aesthetic sensitivity would rate themselves as 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.

vi. 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, namely the "Neue Nationalgalerie", the "Bauhaus-Archiv", the "Hamburger Bahnhof", and the "Nationalgalerie". 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 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).

vii. 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 of the living space. 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 negative correlation was expected.

viii. 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 ($\alpha = .75$) and construct validity have been demonstrated for this scale (Stöber, 1999, 2001).

d. Results

i. Convergent validity

One goal was to examine how the present scale is related to the Test of Aesthetic Judgment Ability. As can be seen in Table 14 (Appendix A), no correlation between scores on the overall aesthetic sensitivity scale and scores on the Test of Aesthetic Judgment Ability

was found (r = .10, p = .30). Similarly, no significant correlations between the scores on the three subscales and scores on the Test of Aesthetic Judgment Ability were found.

Looking at the relationship between the aesthetic sensitivity scale and the CVPA measure shows that there is again no significant correlation between the measures (see Table 14, Appendix A). However, an examination of the correlations with the three different dimensions of the CVPA measure showed significant correlations between the 'response' dimension on the CVPA and the present scale (r = .20, p = .03). This suggests that there is some overlap in the aesthetic sensitivity as assessed by the response dimension of the CVPA and the total score of the aesthetic sensitivity scale. Examining each subscale of the present scale in relation to the CVPA, the space subscale correlated significantly (r =.21, p = .02) with the overall CVPA measure and with the response subscale of the CVPA measure (r =.30, p = .00). The painting subscale showed a significant negative correlation with the values subscale of the CVPA (r =-.23, p = .01). The correlation between the painting subscale and the overall CVPA measure was nearly significant and also negative (r = -.18, p = .05). No significant correlations were found between the cutlery subscale and the overall CVPA measure or its subscales.

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 between the SOP scale and the aesthetic sensitivity scale emerged (see Table 14, Appendix A). Participants with high scores on the visual dimension of the SOP scale also scored higher on aesthetic sensitivity as measured with the present scale (r = .23, p = .01). No significant correlations were found between the subscales of the present scale and the SOP scale. However, for the space and the cutlery subscales the correlations were close to significance (space: r = .17, p = .06; cutlery: r =.18, p = .05).

Next, the present scale's associations with the self-report measures were examined. The statement about judgment certainty was reverse-coded for the analysis. Table 14 (Appendix A) shows that the scores of the overall aesthetic sensitivity scale are not significantly correlated with any of the self-report measures. Nevertheless, a significant negative correlation was found between the self-rating for aesthetic sensitivity and the painting subscale (r = -.22, p = .02). Also, a significant negative correlation emerged between the cutlery subscale and judgment certainty (r = -.26, p = .00).

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 14 (Appendix A) shows no significant correlation between visiting art museum scores and the total score of the aesthetic sensitivity scale (r = .06, p = .53) or its subscales (space: r =.03, p = .79; paintings: r =.07, p = .47; cutlery r =.04, p = .65).

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 spaces. 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 14 (Appendix A) shows that again no significant correlations of the indices with the aesthetic sensitivity scale or its subscales were found.

iii. Social desirability and socio-demographic characteristics

Finally, the scales' relationships to social desirable responding were examined (see Table 14, Appendix A). Scores on the overall aesthetic sensitivity scale (r = -.02, p = .82) and on the subscales (space: r = -.06, p = .54; painting: r =.05, p = .61; cutlery r = -.01, p = .95) were not significantly related to the tendency to respond in a socially desirable way. Scores on the overall scale and its subscales also did not significantly correlate with socio-demographic characteristics such as age, gender and income (all p > .10).

Together, these results show little support for the expected pattern of associations. For the total scale score, convergent validity was found only with the response dimension of the CVPA measure and the visual dimension of the SOP scale. For the three subscales, significant correlations for the space subscale were found with the overall CVPA measure as well as with the response subscale. For the painting subscale significant correlations were found with the value subscale of the CVPA measure and the self-rating in aesthetic sensitivity, yet, unexpectedly, both correlations were negative. Concerning the SOP scale, correlations which were close to significance were found with the space and the cutlery dimensions. Finally, for the cutlery subscale a significant correlation was found with judgment certainty, but again, this correlation was unexpectedly negative. Accordingly, the present study provides little evidence for validity of the present scale.

In order to identify reasons for the surprising lack of convergent and divergent validity, the data of Study 7 and 8 were inspected again, this time with a focus on possible validity problems due to the response format used in the scale.

e. Problems with the Initial Data

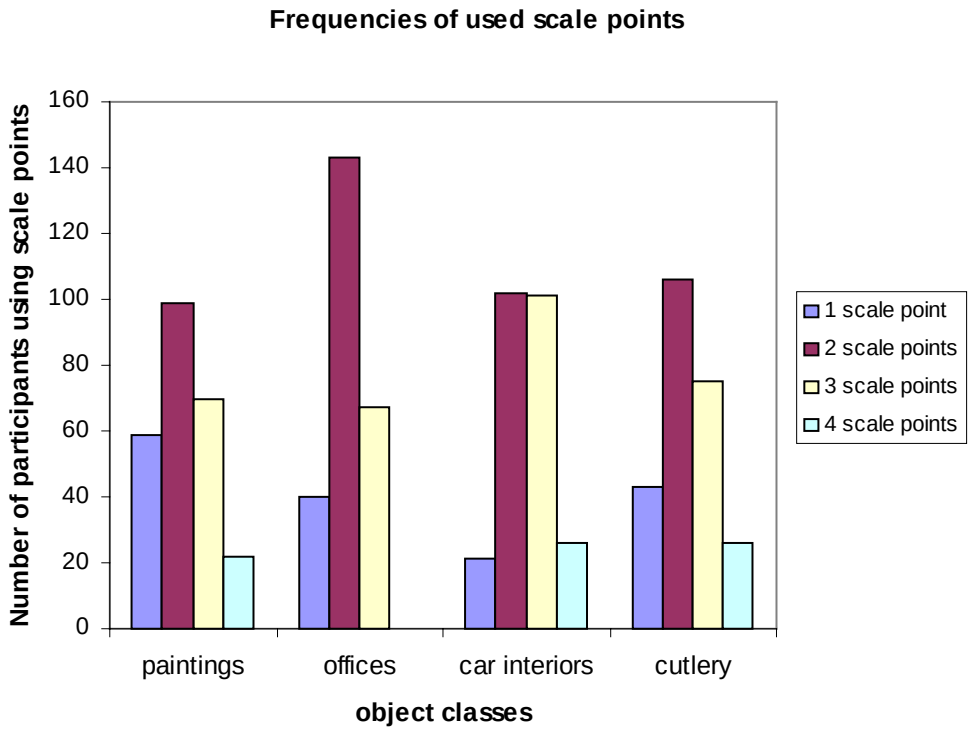
i. Inter-item correlations in Studies 7 and 8

One indication that the results of the present studies might be caused by problems in the data structure are the relatively high inter-item correlation coefficients found in both studies. In Study 7 the inter-item correlations for the three factors are ranging between $\alpha = .84$ and $\alpha = .91$. Calculating the inter-item correlations separately for the office and the car interior stimuli showed similarly high values with $\alpha = .84$ for the office stimuli and $\alpha = .89$ for the car interior stimuli. In Study 8 the inter-item correlations are overall a little lower ranging from .63 to .78 for the three factors. These results were initially interpreted as indications of a high internal consistency of the scale. However, high inter-item correlations within the object classes might also indicate that participants did not use the provided rating scales as expected. Participants were expected to use the rating scales to differentiate between the aesthetic values of the given stimuli. More precisely, they were expected to differentiate between the aesthetic values of the stimuli of different object classes but also within each object class. For maximal differentiation between the stimuli within an object class, participants would need to use as many different points on the rating scales as objects are representing the object class. For instance for the painting stimuli, four different points on the rating scale must be used when judging the aesthetic values of the four different paintings if the person were to make a clear differentiation between all objects of the class. However, if the range of used scale points is much smaller than the number of stimuli in the respective class, then this indicates that respondents did not maximally differentiate between the aesthetic values of the objects. To investigate if the range of provided rating scales was used not only to differentiate between stimuli of different object classes but also between stimuli of the same object class, the range of the scale points used of the provided rating scale was inspected on the individual data level. These (descriptive) analyses are described below.

ii. The range of used scale points

To investigate whether participants differentiated not only between objects of different object classes but also between objects within a class, the data from Study 7 and Study 8 were analyzed on the individual level. More precisely, the range of points each person used on the 7-point rating scale was examined. In Study 7, across participants the entire range of scale points (0 – 6) was used to rate each stimulus. However, an inspection on the individual data level showed that 17 participants used a range of only three different scale points to rate all 15 stimuli. Moreover, three individuals even used only two different scale points to rate the 15 stimuli. The data were then analyzed separately for each object class. Graph 1 shows the number of participants using the seven options of the rating scales to judge the four different stimuli of the painting, car interior and cutlery stimuli (three for offices).

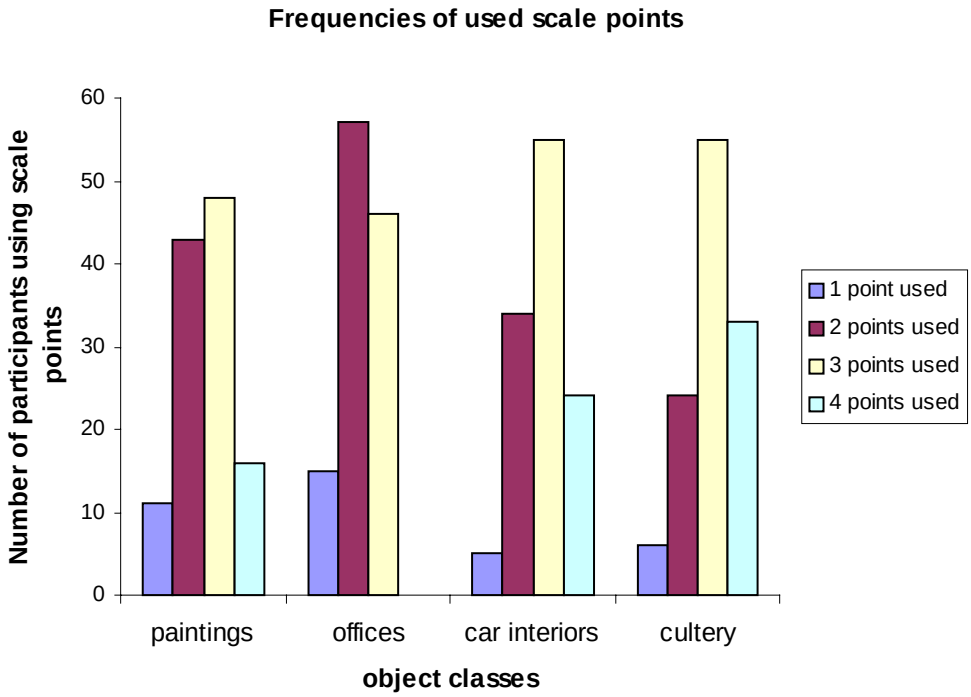
Graph 1. Number of participants using provided scale points of rating scale for Study 7



An inspection of the data of Study 8 showed similar results. Five participants used only a range of three different scale points to rate all 15 stimuli. The data were again analyzed in terms of the range of points each person used on the 7-point rating scale. Overall the entire range of scale points (0-6) was used for 13 of the 15 stimuli. For two of the cutlery stimuli scale points between zero (0) and five (5) were used to judge the aesthetic values of these

stimuli. Graph 2 shows the number of participants who use the different scale points of the rating scale to judge the four, respectively three, different stimuli of each object class.

Graph 2. Number of participants using provided scale points of rating scale for Study 8



In sum, even though the entire range of the 7-point rating scale was used by participants for judging the aesthetic value of the objects, most participants did not maximally differentiate between the objects of an object class. For example, in Study 7 for the three object classes containing four different stimuli, only between 8.8% and 10.4% of the participants used four different scale points on the rating scales to judge the aesthetic values of the given stimuli. For the three office stimuli 26.8% used three different scale points in Study 7. In Study 8 between 13.56% and 27.79% of the participants used four different points on the rating scale to judge the aesthetic values of the stimuli, 38.98% used three different scale points to judge the aesthetic values of the three different office stimuli. In other words, it seems that participants assigned similar values to objects of the same object class and different values to objects of different object classes. In other words, they differentiate between object classes, but not between objects within a specific object class. This tendency seems to be stronger in the North American sample (Study 7), but is also found in the German student sample (Study 8).

These results illustrate that participants did not use the rating scales for differentiating between stimuli within the object classes in both studies. In addition, because the stimuli of all four object classes were presented together, participants probably focused in their comparative evaluations on contrasting between object classes rather than between the four (three) different stimuli within an object class. Consequently, the high inter-item correlations found in the studies are indications of these two problems rather than indications of good internal consistency of the scale. For assessing an individual's aesthetic sensitivity with the present scale, individuals have to differentiate between the stimuli within each object class. More precisely, only if individuals evaluate the aesthetic value of each given stimulus such that this evaluation can be compared to the external standard, then the scale can be said to be a valid measure of aesthetic sensitivity. Because most participants did not differentiate the stimuli within the object classes and thus did not use the rating scale to evaluate the aesthetic value of each stimulus, the scale cannot be said to have assessed their aesthetic sensitivity. Instead it apparently assessed the participants' ability to differentiate between the aesthetic value of the four object classes. Because the object classes were not pre-selected to represent different aesthetic values (e.g., it was not assumed or intended that car interiors overall would be more aesthetic than office interiors), this differentiation between object classes does not measure aesthetic sensitivity. Thus, the low convergent and divergent validity found in Study 8 are most likely the consequence of the measurement approach used (i.e. (1) having respondents evaluate each stimuli separately on a rating scale and (2) presenting the stimuli from all four object classes mixed, instead of separately by object class) rather than indications of the validity of the aesthetic sensitivity scale.

6) Discussion

The aim of the present research was to describe the development and psychometric properties of a scale for measuring visual aesthetic sensitivity towards everyday objects that was constructed using systematically varied stimuli. Results of the EFA (Study 7) showed that the scale consists of three factors, one labeled "space," the second labeled "painting," and the third labeled "cutlery". The test-retest results showed that the scores received with the scale are reliable in terms of temporal stability. Results of Study 8 however did not succeed at providing evidence for the convergent and divergent validity of the scale.

a. Data problem and Choice of Different Methods

When recruiting participants in downtown Montreal it seemed appropriate not to ask them to rank-order the stimuli from the most to the least aesthetic. First of all, for asking participants to rank-order the stimuli, providing them with cards (as done in Study 8) would have been a good method. However, because generally no table was available to order cards, this approach did not seem feasible. More importantly, rank-ordering stimuli is cognitively more demanding for participants than using rating scales. Because participants were not paid for their participation, and because they were approached while going about their daily activities, it seemed the most prudent approach to choose an easy and little demanding assessment method. Consequently, in the paper-pencil version of the scale that they completed they were given 7-point Likert-type rating scales for judging the aesthetic values of the stimuli. This was done with the expectation that participants would use the rating scales to differentiate not only between stimuli of different object classes but also to differentiate the aesthetic values of stimuli within each object class. To investigate if this expectation was met, the data from both studies were analyzed separately for each object class. These analyses showed that even though overall all scale points of the provided 7-point rating scales were used, a large number of participants in both studies used only three scale points or less to indicate the aesthetic values of the four different painting, car interior and cutlery stimuli and only two or less scale points to judge the aesthetic values of the three office stimuli. These results indicate that most participants did not take advantage of the available numeric values of the rating scale to differentiate between objects of an object class. Instead they evaluated the aesthetic value of all objects of the same object class rather similarly, and only differentiated between the aesthetic value of different object classes. This response pattern is also revealed in the exploratory factor analysis, which quite clearly led to the identification of three factors, matching the four object classes (with offices and car interiors in one factor – space). Thus, participants differentiated *between* stimuli of different object classes, but not *within*. Consequently, the data collected for evaluating the reliability and validity of the final scale likely reflect an artifact of the assessment approach itself. In future modifications of the scale, the response format should be changed to a format that enables and encourages respondents to maximally differentiate between stimuli of one object class.

Relatedly, to make sure that different stimuli are evaluated differently within each object class, another method for assessing the aesthetic judgments could be chosen. Even though using methods such as rank order or pair comparison will make it more difficult to administer the present scale in diverse settings (e.g. as paper pencil version) it seems to essential for the present scale. Because the aim of the scale is to measure individual differences in aesthetic sensitivity, the important question is if individuals are able to differentiate between stimuli which may have only subtle differences, i.e. between the objects within an object class that have been manipulated to be differentially aesthetic. Thus, using an assessment method such as rank ordering or paired comparison in which individuals are forced to differentiate between given stimuli is crucial.

b. Stimulus Reduction

Another limitation of the present scale development is the process of stimulus reduction. Choosing the stimuli with the highest separation performance is considered an appropriate approach, however, in the present research additional criteria were applied. For instance, it seemed important to retain stimuli such that each dimension that was implemented when constructing the stimuli was represented in the final set of stimuli. At the same time, another aim was to retain as few stimuli as possible. Even though all these are important criteria, it seems prudent to reduce the scale stepwise rather than reducing it before reliability and validity are assessed. Thus, future research might use the entire set of 34 stimuli and test reliability and validity with a larger number of stimuli. Furthermore, because separation performance depends on the used sample, in the future larger samples from different populations should be considered for stimulus reduction.

In sum, the present research provided only limited evidence for the scale's reliability and validity. Reasons for the lack of psychometric quality were identified in the process of stimulus reduction as well as in the chosen response format (i.e. rating scales). Future applications of the scale, and future investigations of its psychometric properties, should use a version of the scale that forces individuals to differentiate between the stimuli within an object class, such as rank orderings or paired comparisons. Furthermore, the entire stimulus pool of 34 stimuli should be used to build a scale measuring visual aesthetic sensitivity with stimulus reduction being performed at a later point, once the psychometric properties of a long version of the scale have been evaluated.

In Chapter 2 I stressed the need for an empirical identification of the relative importance of relevant judgment dimensions for stimuli. Assuming that all relevant dimensions of a stimulus are equally important for the aesthetic judgment, stimuli with higher values across the set of aesthetic attributes (e.g. "low, high, high") would be expected to be preferred over stimuli with lower values across the set of aesthetic attributes (e.g. "high, low, low"). However, this expectation was not met for all dimensions in Study 6 (Chapter 2). As discussed in Chapter 2, a possible explanation for this finding centers on the relative importance that different dimensions of a stimulus might have for the overall aesthetic judgment. If the dimensions of a stimulus are differently important for an aesthetic judgment, the combination of different levels may lead to different preference orders depending on the importance of a certain dimension, such as those found for some stimuli in Study 6. One aim of the research reported in the following chapter was to investigate the relative importance of each aesthetic dimension used for stimulus construction for the overall aesthetic judgment.

Another related aim of the research reported in the next chapter concerns the external standard used for assessing an individual's aesthetic sensitivity. Measures of aesthetic sensitivity have traditionally investigated how much an individual's aesthetic judgment deviates from an external standard. The extent to which a person agrees with the external standard is then seen as indication of the amount of his or her sensitivity to the aesthetic value of the given stimuli. As mentioned before, in past research external standards have mainly been established in three ways: (a) what experts think is most aesthetic or (b) what the average judgment in a reference group considers as most aesthetic or (c) as a combination of both (e.g., Berlyne, 1971; Child, 1962). However, in all three cases it remains unclear what criteria the judges are using, unless the judgment criteria used for rating the aesthetic value of stimuli are explicated by the respective judges themselves. In effect, existing research has a "clarity of judgment criteria" problem. Moreover, the measures using the existing criteria are really measures of *interpersonal agreement*, rather than measures of a person's aesthetic sensitivity - how sensitive a person is to qualities of the objects - because they rely on how much an individual agrees with the average judgment of experts or of other participants. As a solution to this problem, I propose to use an "optimal" rank order of the stimuli as external standard. Specifically, if an optimal rank order of the stimuli used for scale development can be created, this rank order can serve as the external standard to evaluate an individual's aesthetic sensitivity. This optimal rank order would show which dimensions were important for the judgment of different stimuli, thereby resolving "the clarity of judgment criteria"

problem associated with past research, while simultaneously creating an external standard that is not dependent on interpersonal agreement.

In sum, the aim of Chapter 4 is to determine the relative importance of the relevant aesthetic dimensions for the overall aesthetic judgment and to establish an optimal rank order of the given stimuli that can be used as external standard. Because of the way it is being constructed, this external standard is based on specific knowledge about the judgment criteria and therefore rather independent from a specific reference group of judges. The relative independence from a specific group of judges moves one closer to examining how sensitive a person is to *qualities* of the objects, and away from standards of agreement.