The sound of beauty: How complexity determines aesthetic preference

Jeroen Delplanque\textsuperscript{a,b,}\textsuperscript{⁎}, Esther De Loof\textsuperscript{b}, Clio Janssens\textsuperscript{b}, Tom Verguts\textsuperscript{b}

\textsuperscript{a} Vrije Universiteit Brussel, Belgium
\textsuperscript{b} Ghent University, Belgium

\textbf{A R T I C L E I N F O}

**Keywords:**
Aesthetic preference
Complexity
Prediction error

\textbf{A B S T R A C T}

Stimulus complexity is an important determinant of aesthetic preference. An influential idea is that increases in stimulus complexity lead to increased preference up to an optimal point after which preference decreases (inverted-U pattern). However, whereas some studies indeed observed this pattern, most studies instead showed an increased preference for more complexity. One complicating issue is that it remains unclear how to define complexity. To address this, we approached complexity and its relation to aesthetic preference from a predictive coding perspective. Here, low- and high-complexity stimuli would correspond to low and high levels of prediction errors, respectively. We expected participants to prefer stimuli which are neither too easy to predict (low prediction error), nor too difficult (high prediction error). To test this, we presented two sequences of tones on each trial that varied in predictability from highly regular (low prediction error) to completely random (high prediction error), and participants had to indicate which of the two sequences they preferred in a two-interval forced-choice task. The complexity of each tone sequence (amount of prediction error) was estimated using entropy. Results showed that participants tended to choose stimuli with intermediate complexity over those of high or low complexity. This confirms the century-old idea that stimulus complexity has an inverted-U relationship to aesthetic preference.

1. Introduction

Humans seem uniquely driven to produce and appreciate art. Correspondingly, the study of aesthetics has a long tradition, dating back at least to the Ancient Greeks (Konstan, 2014). Despite this long-standing interest, however, the nature of aesthetics and what objects are considered aesthetically pleasing or beautiful, has remained obscure. One key idea is that to be aesthetically preferred an object should neither be too simple, nor too complex. An object that is too simple would lead to boredom, while too much complexity would cause distraction. The optimal point between the two extremes came to be called the aesthetic middle. This concept had a profound impact on the empirical study of aesthetics. For example, Birkhoff (1933) based his formula for an aesthetic measure on this idea, in which he defined beauty as the ratio of an object's order (simplicity) and its complexity.

The idea of the aesthetic middle was further developed by Berlyne (1971), who formalized the relationship between complexity and preference as an inverted-U shape. As the arousal evoked by a stimulus increases, preference increases up to an optimal point, after which it declines (Berlyne, 1971). A number of variables determine this arousal, including stimulus complexity (Berlyne, 1971; Chmiel & Schubert, 2017). A considerable number of studies confirmed the inverted-U hypothesis for both visual art and music (Aitken, 1974; Chmiel & Schubert, 2017; Gordon & Gridley, 2013; Hekkert & van Wieringen, 1990; Madison & Schröder, 2017; Marin, Lampatz, Wandl, & Leder, 2016; Messinger, 1998). However, these studies used subjective measures of complexity, as rated by the same participants who indicated their aesthetic preference for the stimuli, or by an expert panel of artists and scientists.

Non-rating based measures of stimulus complexity have been used in several visual art studies. For example, Tinio and Leder (2009) presented participants with images containing geometric shapes differing in symmetry and stimulus complexity. Complexity of an image was defined as the number of individual elements (e.g., squares, rectangles, triangles) in the image. Their results showed that images containing many elements were preferred over images containing fewer elements; and that symmetrical images were preferred over asymmetrical ones. This argued against an inverted-U relation between complexity and beauty. However, symmetry and complexity are not independent. Indeed, as symmetry is slightly broken, this leads to increased perceived complexity (Gartus & Leder, 2013). This suggests that complexity cannot just be defined as a function of the number of elements.

More recently, researchers have turned to information theory to define complexity (Schmidhuber, 2009). One useful information-
Theoretical metric is redundancy reduction. Algorithms that detect redundancy have been used for some time to compress the size of computer files (e.g., image compression; Ewert, Dembski, & Marks, 2015; Feldman & Crutchfield, 1998). In one study, image compressibility was considered to define complexity (less compressible is more complex). Participants scored their preference for stimuli at each level of complexity. Again, an increasing effect of stimulus complexity was observed (Forsythe, Nadal, Sheehy, Cela-Conde, & Sawey, 2011; Friedenberg and Liby (2016) applied image compressibility to block patterns, and again obtained an increasing effect. However, the latter pattern needs to be interpreted with caution. Indeed, in addition to redundancy reduction, several other (information-theoretic) measures of complexity were defined, such as pattern density and edge length. Application of these other measures of complexity to the data of Friedenberg and Liby led to either no effect or an inverted-U effect of complexity (Friedenberg & Liby, 2016; Gauvrit, Soler-Toscano, & Guida, 2017). Finally, some studies have demonstrated an inverted-U by applying natural scene statistics, another objective correlate of complexity, using either compressibility (Street, Forsythe, Reilly, Taylor, & Helmy, 2016), fractal dimensions (Spehar, Clifford, Newell, & Taylor, 2003), or the spectral power density slope (Beauvois, 2007; Spehar et al., 2015; Spehar, Walker, & Taylor, 2016).

It would seem, then, that the empirical evidence for an inverted-U relation between complexity and aesthetic preference as proposed by Berlyne (1971) is variable. However, there exists some diversity in both the theoretical and methodological approaches; perhaps as a result of this diversity, results differ across studies, stimuli, and complexity measures (Gauvrit et al., 2017). Second, stimulus preference is typically estimated using Likert scales. Such scales may be sensitive to response biases (Culpepper, Zhao, & Ballenger, 2008; Friborg, Martinussen, & Rosenvinge, 2006; Hartley, 2014), which then requires post-hoc bias control of the scale.

To address the first issue, we consider the effect of complexity from a predictive coding theoretical perspective (Friston & Kiebel, 2009; Van de Cruys et al., 2014; Van de Cruys & Wagemans, 2011). According to this framework, a cognitive system is always actively predicting upcoming stimuli; and unpredicted stimuli lead to more prediction error. Consistently, unpredicted stimuli activate stimulus-specific cortical processing areas more than predicted ones (Egger, Monti, & Summerfield, 2010). Thus, complexity can be defined in terms of prediction error. A stimulus is more complex if its elements are more difficult to predict, leading to more prediction error. Consequently, we will use stimulus material where prediction error is relatively easy to control. This led us to use auditory stimuli, which naturally can be presented in a sequential (and controllably predictable) fashion. Entropy was used as our proxy measure of prediction error (see Method section).

To address the second issue, we estimated aesthetic preference using a two-interval forced-choice (2IFC) paradigm. Forced-choice items are preferable to bias-controlled Likert scales in terms of construct validity (Brown, Inceoglu, & Lin, 2017). In our experiment, subjects were presented with two consecutive tone sequences on each trial, and were asked to indicate which one they prefer. To differentiate between a monotonically increasing or inverted-U relationship between complexity and preference, we tested both linear and quadratic effects of complexity on preference.

2. Method

2.1. Participants

20 participants (18 female; age range 20–35) were recruited using an online platform at Ghent University. Participants attended two sessions. In a first session EEG data was collected; this data will not be considered in this paper. The second session was used for the collection of the behavioural data that are relevant here. Informed consents were signed and collected at the start of the first session. A payment of 35 euros followed the end of the second session. One participant was removed from behavioural data analysis due to an error in data registration.

2.2. Materials

We constructed two baseline sequences of tones (events). Each baseline consisted of a sequence of seven tones played on a grand piano. By deviating from the tones in the baseline, we varied prediction error (entropy) across stimuli. All deviations from the baseline were sampled from the same set of seven tones used to construct the baselines.

The first baseline was called the Flat Baseline. It consists of seven identical tones (B3). This allows us to produce compositions ranging from very repetitive (no deviations) to very chaotic (6–7 deviations). This baseline serves to create the most predictable sequence possible. The second baseline is called the Scale Baseline. This sequence consists of seven different tones from a tonal scale (F3, G3, A3, B3, C4, D4, E4), in which each tone is of slightly higher frequency than the one before (average difference 25.8 Hz). This baseline allows us to capture predictable transitions between non-identical tones. As this is the first study of its kind, we opted for the simplest possible baselines (i.e., one with no deviations and one with a constant increase). However, other baselines could be considered in follow-up research.

Deviations on these baselines are created by first randomly selecting one of the seven tones in the sequence. This tone is then changed by (randomly) replacing it with one of the other six tones of the tonal scale. This process is applied on zero to seven positions in the sequence. Hence, eight deviation categories (0–7) are created.

2.3. Entropy measurement

Even with sequential stimuli, it is hard to estimate prediction error exactly. A useful proxy however, is entropy, which we will use as our measure of complexity. Entropy measures the degree of disorder in a stimulus (or in a collection of events more broadly) (Shannon, 1948). Formally, entropy is defined as follows:

$$Entropy = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

In this formula, a stimulus is comprised of $n$ events (e.g., tones in a composition, colours in a painting, words in a poem), each represented by $x_i$, with $p(x_i)$ being the probability of that event occurring. As an example, when only one event $x_i$ occurs in the stimulus, its probability equals 1, certainty is maximal and entropy equals 0. Entropy measures the deviation from such a perfectly predictable distribution and thus approximates total prediction error.

We define three types of entropy. The first is based on the number of times each unique tone appears in the sequence. For example, in the sequence “F3, F3, G3, A3, C4, D4, E4” tone F3 has a probability of 2/7. Using these probabilities, we calculate first-order entropy; a measure of how complex the sequence is in terms of tone repetitions. Fig. 1a and b illustrates the calculation of first-order entropy on the two baselines (more examples appear in Table 1). Second, we looked at the transitions from one tone to the next. Here, we counted how many positions a tone was higher or lower on the tonal scale than the one that preceded it. The entropy based on the probability of these transitions is referred to as second-order entropy. As an example, in the sequence “F3, F3, G3, A3, C4, D4, E4” the transition of the first to the second tone has a positional difference of zero, since both are identical. However, the transition of the second to the third tone has a difference of 1. In the example, one out of six transitions has a difference of 0, four out of six have an (absolute) difference of 1, while one out of six has a difference of 2. See Fig. 1c and d for an illustration of how second-order entropy is calculated on the two baselines. Finally, we calculated an average
the trial. Afterwards, a question appeared on the screen asking the participants whether they heard a difference between the two sequences. Participants had to press left (Q) on a keyboard if they did not hear a difference and right (P) if they did. All sequences in a pair were actually different from each other, but preference judgments are meaningless if the participants failed to notice the difference. Therefore, trials in which the difference was not noticed were excluded from further analysis. This was followed by a second question, checking which sequence they preferred. Participants had to press left if they preferred the first and right otherwise. Both questions remained on the screen for up to 5000 ms, or until a response was registered. The trial ended with an inter-trial interval (a blank screen of 1000 ms).

For every deviation category (0–7), we selected two sequences per participant, one based on the Flat Baseline and one on the Scale Baseline. Hence, every participant was presented with a different randomly generated set of sixteen sequences. Every possible comparison between all sixteen sequences was presented, without allowing repetitions. However, due to time constraints, not every sequence appeared in both the first and second position. Instead, the order of appearance in a sequence pair was randomized across participants. In total there were 120 trials, each with a unique pair of sequences, and each deviation category appeared 30 times.

### 2.5. Statistical analysis

We defined preference as the choice between the first and second stimulus in the 2IFC task. Both a linear and a quadratic effect of entropy on preference was tested using a linear mixed effects model (LME), the results of which are reported using a type III Wald chi square test. The quadratic components specifically tested an inverted-U relationship between entropy and stimulus preference. To calculate the quadratic component, the average entropy was subtracted from each entropy value and these centred entropy values were then squared. As two stimuli are presented on each trial, there are two linear components and two quadratic components; one for the first stimulus and one for the second. A separate analysis was performed for each the three types of entropy. In the analysis we only included trials in which the participant heard a difference between the first and second stimulus.

### 3. Results

The first sequence was preferred on 44% of trials. On most trials the participants heard a difference between the two sequences (92%). First-order entropy of the first stimulus had a significant effect on preference, with both its linear, $b = 0.53, \chi^2(1, N = 19) = 19.5, p < 0.001$, and quadratic, $b = -0.16, \chi^2(1, N = 19) = 11.6, p < 0.001$, components being significant (Fig. 3a, column 1). Similar results were found for the second stimulus in the task with its linear, $b = 0.77, \chi^2(1, N = 19) = 31.3, p < 0.001$, and quadratic, $b = -0.22, \chi^2(1, N = 19) = 21.0, p < 0.001$, effects (Fig. 3a, column 2).

In the case of second-order entropy a marginally significant linear effect was found for the first stimulus, $b = 0.24, \chi^2(1, N = 19) = 3.84, p = 0.050$, and no quadratic effect, $b = -0.057, \chi^2(1, N = 19) = 1.36, p = 0.24$ (Fig. 3b, column 1). The second-order entropy for the second stimulus on the other hand showed both significant linear, $b = 0.56, \chi^2(1, N = 19) = 16.5, p < 0.001$, and quadratic, $b = -0.50, \chi^2(1, N = 19) = 83.4, p < 0.001$, effects (Fig. 3b, column 2).

The linear effect of average entropy for the first stimulus was not significant, $b = 0.20, \chi^2(1, N = 19) = 1.65, p = 0.20$. The quadratic effect on the other hand was significant, $b = -0.24, \chi^2(1, N = 19) = 31.4, p < 0.001$ (Fig. 3c, column 1). Average entropy for the second stimulus showed significant effects, both linear, $b = 0.70, \chi^2(1, N = 19) = 13.5, p < 0.001$, and quadratic, $b = -0.21, \chi^2(1, N = 19) = 18.8, p < 0.001$ (Fig. 3c, column 2).

### Table 1

Examples of stimuli based on the two baselines with each type of entropy.

<table>
<thead>
<tr>
<th>#</th>
<th>Example stimulus</th>
<th>Flat Baseline</th>
<th>Scale Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>B3 B3 B3 B3 B3 B3</td>
<td>0 0 0 0 0 0</td>
<td>1.330 1.230 1.450 1.390 1.380 1.400</td>
</tr>
<tr>
<td>1</td>
<td>B3 B3 B3 G3 B3 B3</td>
<td>0.410 0.868 0.639</td>
<td>1.090 1.090 1.090</td>
</tr>
<tr>
<td>2</td>
<td>B3 E4 D4 B3 B3 B3</td>
<td>0.796 1.242 1.019</td>
<td>1.090 1.090 1.090</td>
</tr>
<tr>
<td>3</td>
<td>E4 B3 B3 C4 F3 B3</td>
<td>1.154 1.561 1.357</td>
<td>1.090 1.090 1.090</td>
</tr>
<tr>
<td>4</td>
<td>B3 A3 F3 B3 A3 E4</td>
<td>1.277 1.561 1.419</td>
<td>1.090 1.090 1.090</td>
</tr>
<tr>
<td>5</td>
<td>B3 D4 B3 E4 A3 D4 A3</td>
<td>1.352 1.561 1.456</td>
<td>1.090 1.090 1.090</td>
</tr>
<tr>
<td>6</td>
<td>E4 D4 B3 F4 C4 D4</td>
<td>1.550 1.792 1.671</td>
<td>1.090 1.090 1.090</td>
</tr>
<tr>
<td>7</td>
<td>D4 C4 A3 G4 E4 C4</td>
<td>1.550 1.330 1.440</td>
<td>1.090 1.090 1.090</td>
</tr>
</tbody>
</table>

Note: the # sign above the first column indicates the number of deviations. Numbers in columns 3–5 are entropy values.
Based on predictive coding theory, we applied the concept of entropy from information theory to measure the complexity of auditory sequences. We empirically related this entropy to the aesthetic preference for those sequences. In line with the hypothesis of Berlyne (1971), the inverted-U is confirmed by the quadratic terms in the LME. Because we modelled the probability of choosing the first stimulus, an inverted-U pattern predicts negative and positive regression weights for the first and second stimulus, respectively. This is indeed what we found. In addition to the quadratic term we also found significant linear effects, demonstrating that the inverted-U is not symmetrical. This may be due to having sampled more from the left, than from the right side of the complexity range. However, this would have to be confirmed in future work.

Our results are in line with early studies that used subjective (rating-based) measures of complexity (Aitken, 1974; Hekkert & van Wieringen, 1990; Messinger, 1998) and support several recent studies using a number of objective measures of complexity (Spehar et al., 2015; Spehar et al., 2016; Street et al., 2016). However, our data contradicted other research using other measures of complexity, specifically in visual art (Forsythe et al., 2011; Friedenberg & Liby, 2016; Gauvrit et al., 2017; Tinio & Leder, 2009). Several factors may have contributed to this discrepancy. First, any increasing effect of complexity is not necessarily incompatible with an inverted-U shape (Chmiel & Schubert, 2017). For instance, if one samples only the left part of the complexity range, one would find an increasing effect. Second, we used a 2IFC paradigm rather than asking subjects to rate stimulus preference. Third, our theoretical emphasis on prediction error prompted us to use auditory stimuli, whereas most earlier studies used non-rating based measures of complexity used visual material. Future work is needed to disentangle these possibilities.

Although predictive coding has a long history in psychology (e.g., Rescorla & Wagner, 1972), it has been considerably broadened recently to encompass all aspects of cognition (Bastos et al., 2012; Rao & Ballard, 1999). At the broadest level, predictive coding postulates that a cognitive agent predicts upcoming stimuli and actions. This is useful in both perception and learning, as it allows accurate estimation of parameters attached to these stimuli and actions (e.g., luminance of stimuli, or value of actions). Deviations from these predictions are registered as prediction errors. This concept has turned out to play an important role in, for example, visual perception (Summerfield & Egner, 2009), auditory perception (Bekinschtein et al., 2009), associative learning (den Ouden, Friston, Daw, McIntosh, & Stephan, 2009), declarative memory (De Loof et al., 2018), and the experience of pleasure (Inzlicht, Bartholow, & Hirsh, 2015).

An important role of unexpected events, or prediction errors, has more recently been suggested in aesthetic judgment (Chetverikov & Kristjánsson, 2016; Trapp, Shenhav, Bitzer, & Bar, 2015; Van de Cruys et al., 2014; Van de Cruys & Wagemans, 2011). One recent study observed a higher preference for predictive stimuli relative to non-predictive (i.e., randomly predictive) ones (Braem & Trapp, 2017). This seems consistent with the decreasing trend at the right side of the inverted-U, although we must be cautious: First, this work found an effect specifically for predictive, not for predictable, stimuli. Second, their comparisons were binary (predictive versus non-predictive), which makes evaluating the underlying preference curve very hard. Future research will determine whether the preference curve for predictive stimuli also has an inverted-U shape, as for predictable stimuli. Why would prediction errors show an inverted-U effect on aesthetic preference? This may be illustrated with a classical problem from Reinforcement Learning, the $n$-armed bandit (Sutton & Barto, 1998). Here, subjects are faced with a slot machine (bandit) with $n$ handles. The reward delivery probabilities of the handles are unknown to the subject. All existing algorithms to solve this problem face the exploration-exploitation trade-off; they must find a balance between exploiting known (good) arms and exploring potentially more rewarding ones (Cohen, McClure, & Yu, 2007; Sutton & Barto, 1998). Considered in the current context, this balance corresponds to finding an optimal (intermediate) level of prediction error. Indeed, too little prediction error means that one is not sufficiently exploiting; whereas too much prediction error means that one is not sufficiently exploiting. Thus, we propose that subjects prefer an intermediate level of entropy because this leads to an intermediate level of prediction errors, and an optimal balance between exploration and exploitation.

Our work suggests several avenues for future research. First, as just noted in the previous paragraph, future work should directly test whether intermediate levels of prediction error are preferred. For this purpose, one can consider estimating the parameters of the prediction machinery, and thus the prediction errors, as in Howard-Jones, Demetriou, Bogacz, Yoo, and Leonards (2011). Unfortunately, this is not feasible in the current experimental paradigm as there are no
behavioural data that are informative to estimate these parameters. Second, our theoretical emphasis on prediction error prompted us to use auditory stimuli, which are most easily manipulated in their predictability. It is however possible that predictive coding applies best to auditory stimuli. To test this, future studies should apply the current design to visual stimuli. Earlier research suggests similar predictive-coding mechanisms for visual stimuli are possible (Van de Cruys & Wagemans, 2011). Third, there is a need to compare measures of complexity in a systematic manner, including entropy (Gauvrit et al., 2017), fractal dimensions (Spehar et al., 2003), and the spectral power density slope (Beauvois, 2007; Spehar et al., 2015; Spehar et al., 2016). Such investigation may also reveal (in)consistencies between these measures and their associated theoretical frameworks. Fourth, interactions of prediction errors and individual-difference factors must be

Fig. 3. Effect of entropy (x-axis) on the probability (%) of choosing either the first (column 1) or second (column 2) stimulus. Panels a, b, and c for first-order, second-order, and average entropy, respectively. The regression bands indicate the 95% confidence interval.
studied. Indeed, from a predictive coding perspective, all stimulus processing is inherently of a predictive and therefore subjective nature. For example, a musical piece that is predictable for one subject, may not be predictable for another. Thus, different subjects would have their maximum preference at different levels of stimulus complexity. The role of familiarity can also be explained and studied using the predictive coding framework. Studies have shown that the inverted-U can change depending on a subject’s familiarity with the stimulus (Chmiel & Schubert, 2017; Hunter, Schellenberg, & Schimmack, 2008). This fits well with a predictive coding account of aesthetic preference, as a system improves in predicting a stimulus with repeated presentations. Other individual variables such as gender (Marin & Leder, 2013) and personality (Hunter et al., 2008) can also influence the inverted-U. These interactions are yet to be established with a non-rating based measure of complexity. In general, we believe that bridging the gap between current cognitive theory and empirical aesthetics will be an important step in the psychological study of beauty.

Acknowledgements

TV was supported by BOF17-GOA-004 grant from Ghent University. We thank Sennes Braem for useful comments on a previous draft of this paper. Correspondence about this paper can be sent to Jeroen Delplanque (jeroen.delplanque@vub.ac.be) or to Tom Verguts (tom.verguts@ugen.t).

