Continuous aesthetic judgment of image sequences

Mel W. Khaw¹,²,⁎, David Freedberg¹,³

¹ Italian Academy for Advanced Studies in America, Columbia University, United States
² Department of Economics, Columbia University, United States
³ Department of Art History and Archaeology, Columbia University, United States

ARTICLE INFO

Keywords: Aesthetics
Sequence effects
Perception
Judgment
Decision making

ABSTRACT

Perceptual judgments are said to be reference-dependent as they change on the basis of recent experiences. Here we quantify sequence effects within two types of aesthetic judgments: (i) individual ratings of single images (during self-paced trials) and (ii) continuous ratings of image sequences. As in the case of known contrast effects, trial-by-trial aesthetic responses are negatively correlated with judgments made toward the preceding image. During continuous judgment, a different type of bias is observed. The onset of change within a sequence introduces a persistent increase in ratings (relative to when the same images are judged in isolation). Furthermore, subjects indicate adjustment patterns and choices that selectively favor sequences that are rich in change. Sequence effects in aesthetic judgments thus differ greatly depending on the continuity and arrangement of presented stimuli. The effects highlighted here are important in understanding sustained aesthetic responses over time, such as those elicited during choreographic and musical arrangements. In contrast, standard measurements of aesthetic responses (over trials) may represent a series of distinct aesthetic experiences (e.g., viewing artworks in a museum).

Judgments of absolute magnitude (e.g., of brightness, loudness, etc.) as well as subjective valuation (e.g., of monetary lotteries) appear to be made with respect to changeable reference points (De Martino, Kumaran, Holt, & Dolan, 2009; Stewart, Brown, & Chater, 2005). As a consequence of this reference dependence, such judgments change on the basis of contextual factors such as presentation order (Hellström, 2001) and preceding judgments (Stewart, Brown, & Chater, 2002). To what extent do these principles apply to aesthetic valuation? Sequential effects arising from recently viewed stimuli have been shown to affect the perceived beauty of photographs (Tousignant & Bodner, 2014), and facial attractiveness (Chang, Kim, & Cho, 2017; Cogan, Parker, & Zellner, 2013; Pegors, Mattar, Bryan, & Epstein, 2015). A common effect found in such studies is a stimulus contrast effect – ratings are biased away from the visual pleasantness of the previously presented image. In addition, extensive familiarization toward either complexity or simplicity results in increased aesthetic preferences for the contrasting attribute (Tinio & Leder, 2009). Indeed, virtually all influential theories of aesthetic experience strive to accommodate past experiences in some way, whether in the form of novelty (Berlyne, 1971), prototypicality (Martindale, Moore, & West, 1988), or memory integration (Leder, Belke, Oeberst, & Augustin, 2004). However, the link between aesthetic responses measured over single trials and those experienced over longer timescales remain unknown. Here we shed further light on sequential aesthetic processing by studying judgments made toward discrete and continuous presentations of images.

At present, we aim to compare aesthetic judgments made across two modalities: ratings of single images and continuous ratings provided over image sequences. This will allow us to test whether sequences elicit identical aesthetic responses as their constituent images when judged in isolation. Trial-by-trial responses to individual images remain the standard level of analysis for studies of aesthetic judgments (for a review, see: Leder & Nadal, 2014). On the other hand, the continuous methodology is an extension of the method used by Kahneman, Wakker, and Sarin (1997) to observe regularities in subjective experiences over time. Kahneman et al. (1997) were originally interested in the momentary affect of subjects viewing video clips. Variants of this method were also used to track momentary hedonic enjoyment (Kahneman, Kahneman, & Tversky, 2003) as well as discomfort (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993). We introduce this method as a means of measuring aesthetic value over a fixed interval of time (during the presentation of a sequence of images). By assembling images into sequences, we hope to reproduce (in a simplified manner) effects present during the continuous and structured nature of artistic delivery (such as in songs, plays, film, etc.). In contrast, trial-by-trial presentations may be seen as analogous to viewing separate artworks in a gallery featuring a mixture of themes.
represent a unique stimulus class that can be presented without alteration in both discrete and continuous formats. Thus, the use of images as stimuli bears two main advantages, namely: (i) images are presentable identically for both measures; and on this basis, (ii) average responses across both cases are comparable. Image sequences also presentable identically for both measures; and on this basis, (ii) average images as stimuli bears two main advantages, namely: (i) images are iteration in both discrete and continuous formats. Thus, the use of represent a unique stimulus class that can be presented without a particular order of presentation.

Given the use of specific choreographic phrases (Blom & Chaplin, 1982) and song structures (Lomax, 1962), as well as individuals’ capacity for narrative-based reasoning (Cohn, Jackendoff, Holcomb, & Kuperberg, 2014), we aim to test for arrangement-level differences in aesthetic response. We hypothesize that single and sequential presentations will yield different aesthetic ratings as well as underlying sequence effects. First, using discrete (trial-by-trial) measures of aesthetic judgments, we test whether such judgments exhibit previously observed sequence effects. Next, we record moment-to-moment judgments while participants observe image sequences (constructed using a subset of the same images). This allows us to compare average responses and relevant adjustment dynamics across the two modes of presentation. In addition, we test for declarable preferences toward particular sequence types using a binary choice task.

1. Method

1.1. Participants

26 adults (16 female; $M = 23.54$ years, $SD = 3.02$ years, range = 18–28 years) participated in all three phases of the study consecutively. Participants were Columbia University undergraduate and graduate students recruited from advertisements placed throughout the university’s main campus. Participants completed each task at a private computer station in about 60 min. Participants were briefed about the details of all three tasks and were compensated with $10 for participating. All procedures were approved by the Columbia University Institutional Review Board (protocol #IRB-AAAR3901).

1.2. Procedure

We began our experiment by measuring aesthetic ratings of individual stimuli presented in pseudorandom order. Then, subjects faced two additional tasks. In one of these tasks we presented a series of image triplets in a continuously changing presentation while recording momentary aesthetic ratings in real time. In effect, we first measure trial-by-trial ratings of images drawn from alternating categories (e.g., a dance posture, followed by a fractal pattern). Subsequently, the continuous condition presents triplets of images drawn from a single category (e.g., three dance postures). We also conducted a choice experiment with a subset of the constructed triplets to test whether preferences are spontaneously declarable (i.e., without having to view them sequentially) for different sequence types. The order of the continuous rating and choice tasks were counterbalanced across participants in order to control for familiarity and confirmation bias effects.

1.2.1. Experimental stimuli

Each phase of the experiment described below utilizes a single set of 36 image stimuli. This set included both artificially-generated fractal patterns and illustrated human poses (Appendix Fig. 1). Of the human pose images, stimuli consisted of solo (ballet) and partner (swing) dance poses. Abstract stimuli were generated using an automated fractal generator (XaoS) and human poses were comprised of free license vector images. Human poses and fractal patterns were used following the basic assumption that they were likely to elicit non-zero aesthetic ratings (based on their previous applications in the extant literature). All images were rendered in grayscale for uniformity.

![Image 1](image-url)

**Fig. 1.** (a) First, subjects responded with individual ratings from a scale of 0–10 on stimuli presented in random order. Stimuli were drawn from different categories consecutively, so as not to induce perceptions of continuity (e.g., a ballet image could only be followed by a fractal or partner dance image). (b) In the continuous rating phase, subjects’ ratings were continuously recorded as they were shown triplets from single categories (each individual image was presented for 3 s). (c) In the choice phase, subjects chose between repeating or non-repeating sequences which triplet they would prefer to witness sequentially.

1.2.2. Discrete rating phase

In the discrete rating phase (Fig. 1a), subjects were presented with members of each stimulus category in pseudorandom order – consecutive stimuli were chosen to be from different categories but were otherwise chosen randomly from their respective sets. Subjects were instructed to rate on a continuous slider bar how pleasing they perceived the on-screen stimuli, ranging from 0 (not pleasing) to 10 (most pleasing). Subjects were able to declare their judgments up to a precision of two decimal places and were not limited in the time available to declare their ratings. Each image was presented with three repetitions, amounting to 108 trials for 36 image stimuli. The reliability and variability associated with each image’s ratings, along with details of these measures, are available in Appendix Fig. 2.

1.2.3. Continuous rating phase

In the continuous rating phase (Fig. 1b), participants were presented with sequences of triplets consisting of members of a single category (e.g., three fractal images). Each member of the triplet was presented for 3 s while subjects continuously indicated their rating (on the same scale as the initial discrete rating phase). In order to separate the presentation of each sequence, an inter-trial interval of 1–3 s (randomly
determined on each trial) was implemented, during which participants were presented with the message: “Preparing next sequence.”

Unlike the randomized nature of the discrete trial arrangement, care was taken to construct all possible combinations of triplets using stimuli at low, medium, and high ranges of a participant’s rating distribution. This was to ensure that sequences did not introduce drastic fluctuations in aesthetic value across images. These ranges were based on average ratings of each image provided by each subject during the discrete rating phase. To this end, three sets of exemplars from each category were chosen from the following range of values: below the 25th percentile (“low” ratings), between the 25th and 75th percentile (“medium” ratings), and above the 75th percentile (“high” ratings). Triplets were then constructed with replacement, yielding 27 unique triplets. Testing 27 triplets in each range results in a total of 81 continuous rating trials. Importantly, any assembled triplet can be classified as either entirely repeating (i.e., AAA) or not entirely repeating (e.g., ABA, AAB, etc.).

1.2.4. Choice phase

In the choice phase (Fig. 1c), participants were presented with binary choices between entirely repeating (AAA) and not entirely repeating (e.g., AAB, ABA, etc.) sequences. Sequences were displayed as panels consisting of three images to depict a triplet set in its entirety. Thus, entire sequences were presented simultaneously as a “sequence plan” consisting of three panels each. Participants were instructed to select which “sequence plan” they would prefer to view as consecutive images (i.e., hypothetically viewing each image sequentially from top panel to bottom panel). 30 pairwise choices were presented from each images (i.e., hypothetically viewing each image sequentially from top panel to bottom panel). 30 pairwise choices were presented from each images (i.e., hypothetically viewing each image sequentially from top panel to bottom panel). 30 pairwise choices were presented from each images (i.e., hypothetically viewing each image sequentially from top panel to bottom panel).

2. Results

2.1. Discrete rating results

We first test for sequence effects in the traditional psychophysical sense. A simple way to identify sequence effects across trials is to perform a linear regression on responses using previous stimulus values (e.g., objective stimuli magnitudes) and/or participant responses (Jesteadt & Wier, 1977). A similar regression approach was implemented by Pegors et al. (2015) on preceding hair color and facial attractiveness ratings. In the present study, the following linear regression equation was fitted to each subjects’ ratings:

\[ r_i = \hat{\beta}_0 + \sum_{n=1}^{i} \hat{\beta}_n r_{i-n} + \epsilon \]

where \( r_i \) refers to the rating produced on trial \( i \) or \( i - n \) trials prior. Following Pegors et al. (2015), we verify an acceptable degree of multiple collinearity in our regressors by computing the Variance Inflation Factor (VIF). The average VIF across regressors are low \( (r_{-1} = 1.12, r_{-2} = 1.22, r_{-3} = 1.12) \), with serious levels of multiple collinearity usually regarded as values above 10 (O’brien, 2007). Each regression was solved using ordinary least squares for each subject, with average and individual parameter estimates plotted in Fig. 2. Using one-sample t-tests on the distribution of each parameter estimate, we confirm a significant contrast effect (negative regression weight) arising from the preceding trial’s rating \( (t(25) = -5.55, p < 0.001) \).

2.2. Continuous rating results

2.2.1. Continuous versus discrete ratings

First, we sought to determine to what extent continuous ratings differ from the independent elicitation of ratings across trials. To quantify this relationship, we identify the peak (maximum) rating in the continuous phase and their corresponding discrete ratings (on average for each image). Identified peaks (maximum ratings recorded during an image’s presentation) are correlated strongly with average ratings in the discrete phase \( (r = 0.75, p < 0.001) \). Peak continuous ratings are thus related but not perfectly predicted by their discrete counterparts.

2.2.2. Arrangement-related differences in ratings

Next, we ask whether different sequences elicited different rating behavior conditional on their presentation order. As no analytical precedent has been set for a task of this kind, we compared simple averages of three basic metrics: (i) ratings measured at all moments of presentation; (ii) adjustments made during presentation; and (iii) differences in ratings between discrete and continuous presentations. Under the null hypothesis of sequence-invariant aesthetic response, averages on these metrics should be uniformly distributed between sequences.

The first metric involves comparing the average ratings recorded at every moment for each sequence type: ABA (middle element change), AAB (final element change), and ABB (first element change). For all statistical tests performed on these ratings, our analysis window begins at the onset of the second element’s presentation, as subjects use the initial 3-second window to initialize their ratings from the starting position of zero. ABA sequences elicited the greatest average ratings compared to other sequence types \( (z_{ABA} = 8.40, z_{ABB} = 7.31, z_{AAA} = 6.98; p < 0.001; \) Wilcoxon rank sum tests), although the average magnitude of this difference is modest (Fig. 3b).

As a more precise test of differences between conditions, we ask whether the moment-to-moment ratings depart from the average ratings associated with each image (as obtained in the discrete phase when individual images were rated in a trial-by-trial manner). In order to best equate the continuous recordings to the discrete ratings, we compute these potential differences conditional on the slider bar being stationary. To do this, we first identify momentary fluctuations in continuous ratings across sequences. Unsurprisingly, the ABA sequence type elicited the greatest average magnitude of adjustments \( (z_{ABA} = 10.94, z_{ABB} = 10.05, z_{AAA} = 24.11; p < 0.001; \) Wilcoxon rank sum tests; Fig. 3c). In order to control for such adjustment periods, we performed the subsequent analysis after removing ratings that contained momentary adjustments (only considering ratings at time \( t \) that satisfy the property \( Rating_t = Rating_{t-1} \)). On these remaining stationary ratings, we compute changes between continuous and discrete.
ratings (for a particular image) as:

\[
\% \text{Change}_{i,t} = \frac{r_i - r_t}{r_t} \times 100
\]

where \(i\) refers to the identity of the image stimuli currently being presented and \(t\) refers to the stationary rating recorded during that millisecond interval. \(r_t\) is then the average rating that was declared by the subject during the discrete (trial-by-trial) rating phase. Again, we observe that ABA sequences are associated with the greatest positive increase in ratings (\(z_{ABA} = 12.33, z_{ABB} = 4.12, z_{AAA} = 19.80; p < 0.001;\) Wilcoxon rank sum tests).

2.3. Choice results

Here we test whether subjects prefer entirely repeating (AAA) sequences when asked to choose between these and remaining sequence types (e.g., AAB, ABA, etc.). Under the null hypothesis (that sequences do not affect aesthetic response), subjects should be indifferent toward the presented alternatives, resulting in choice probabilities for either category to be around 50%. This would be due to our procedure of assembling sequences that have approximately equal ratings (on average, for each individual). On average (unconditional on the alternative sequence type), this appears to be the case (Fig. 4, denoted as “Average” condition). However, examining each sequence type reveals a significant preference for ABA sequences, and a preference against AAB and ABB arrangements (Fig. 4). A one-way ANOVA indicates a significant difference between groups, \(F(2, 1715) = 27.47,\)
p < 0.001), confirming an uneven distribution of choices, despite an even presentation of sequence types. Formal comparisons between the fraction of choices made toward ABA sequences compared to AAB or ABB sequences confirm this difference (z_{AAB} = 6.15, z_{ABB} = 6.51; p < 0.001; Wilcoxon rank sum tests).

2.4. Main effects by image category

Finally, we verify whether two main effects are robust across the categories of images used in the study. These phenomena are (1) discrete versus continuous rating differences and (2) choice preferences for ABA sequences. Using two-sided t-tests, the increase in adjusted ratings within ABA sequences is significantly greater than zero for all three categories: t_{t_{ANOVA}} = 152.46, t_{t_{partner}} = 72.33, t_{t_{fractal}} = 28.00, p < 0.001 (Fig. 5a). Secondly, the average preference for ABA sequences is observed for both dance pose categories. In the case of fractal patterns, subjects indicate no significant preference for a particular sequence type. Using two-sided t-tests against the null hypothesis that the fraction of ABA choices is 0.5, t_{t_{ANOVA}} = 4.36, t_{t_{partner}} = 6.58, p < 0.001; while t_{t_{fractal}} = −0.94, p = 0.35 (Fig. 5b).

3. Discussion

To recapitulate our findings, we find that aesthetic judgments carry distinct sequential dependencies across discrete and continuous measurements. In a traditional trial-by-trial elicitation of judgments, current ratings are negatively correlated with ratings given to the preceding image. In a continuous rating paradigm, the onset of change introduces a net increase in ratings (relative to isolated presentations). These increases in ratings appear to be meaningfully related to the aesthetic value derived from viewing such sequences – the same subjects indicate a preference for change-rich sequences in an independent binary choice experiment. Thus, in the continuous domain, sequences featuring any level of variety are able to elicit a greater aesthetic response than isolated presentations of their constituent images.

The contrast effect we observe is a replication of effects previously seen in judgments of facial attractiveness (Cogan et al., 2013; Pegors et al., 2015) and photographic beauty (Tousignant & Bodner, 2014). These results suggest that, on average, aesthetic ratings are biased away from the average rating given to the preceding image. We note that the effect we see in our current study is robust across regressors that specify either the average rating of preceding images or the previous ratings themselves (these variables are highly correlated, r = 0.95). More generally, contrast effects occur in judgments of high-level perceptual judgments such as categories (Stewart et al., 2002), as well as basic intensities such as loudness (Jesteadt, Luce, & Green, 1977) and sweetness (Schiesser & Frijters, 1992). The given interpretation of such contrast effects also apply here. That is, stimuli from the preceding trial appear to act as a reference point for the current judgment. In addition, such negative correlations are consistent with biophysical models of subjective value encoding (Louie, LoFaro, Webb, & Glimcher, 2014), where value-related neural activation is divisively normalized by previously experienced input, resulting in contrasts levels of neural activity.

Another previously observed sequence effect is a positive correlation (i.e., assimilation effect) toward previous responses (Pegors et al., 2015; Petzold & Haubensak, 2001) or mean attractiveness of prior stimuli (Chang et al., 2017). The absence of an assimilation effect here might be due to the alternating stimuli categories across our discrete phase trials. Consecutively presenting images with similar thematic content (e.g., faces) may foster the perception of continuity and strengthen such effects. Under this interpretation, effects observed in our continuous paradigm may be seen as an extreme case of assimilative percepts.

During the viewing of continuous sequences, the occurrence of change is associated with an increase in ratings relative to independent viewings (Fig. 3d). Moreover, this increase in ratings is consistent with an average preference – indicated separately by choice – in favor of ABA sequences. ABA arrangements indeed stand out as having the greatest number of changes relative to other sequences in our experiment. This bias is unlikely to be driven by previously documented contrast effects as sequences were constructed using (approximately) equally-rated images from the entire rating range. In addition, we test and confirm that change-related increases are greater for high-value sequences compared to low-value sequences (Appendix Fig. 3); this suggests that these adjustments occur in the opposite direction of a contrast effect. Thus, there appears to be different dynamics of aesthetic response when judgments are made and varied continuously. As we restricted all elements of sequences to lie within the same percentile range, future work should study how change interacts with different ranges of aesthetic value (e.g., low followed by high-value stimuli) to affect subsequent judgments.

Surprisingly, subjects' choices (but not ratings) indicate that sequences featuring early or late change points (e.g., ABB or AAB sequences) are undesirable relative to AAA sequences. This result can be understood as a chosen preference for symmetric (e.g., AAA, ABA)
arrangements (at least when presented simultaneously), similar to the results presented by Orgs et al. (2013). With this view, the aggregate preference ordering in choices (ABA ≻ AAA ≻ ABB – AAB) is explained by preferences for both variety and symmetry, with ABA types exemplifying both features during a simultaneous presentation of its composite images. Fractal patterns may additionally elicit other kinds of symmetry preferences, as discussed in further detail below.

The sequence-level effects explored here sharpen our understanding of sustained aesthetic response. The impact of arrangement styles has been previously noticed and analyzed by scholars from different fields – relevant observations include repeated structural units in myths (Lévi-Strauss, 1955), and the pervasive nature of AABA song structures in popular music (Fitzgerald, 1996). Indeed, studies in the perception of music have revealed distinct cognitive processing at the structural level (Koelsch, Rohrmeier, Torrecuso, & Jentschke, 2013), with high-level effects of musical properties such as melodic familiarity (Pereira et al., 2011) and expectation (Pearce & Wiggins, 2006). The present study can be seen as a starting point for the study of aesthetic response toward experienced arrangements. This approach complements other studies of continuous aesthetic response, such as those tracking the emergence of perceptual realizations in art (Muth, Raab, & Carbon, 2015, 2016). In addition, the methods proposed here are generalizable to any episodic presentation of stimuli, without being restricted to experiences of bistable perceptions and sudden insights. Future experiments of this kind may include longer durations and dynamic stimuli to explore the interplay between reported visual pleasantness and shifts in participants’ beliefs.

We also begin to address questions of generality using analyses separated by individual image categories: human dance poses (solo and partnered) and fractal patterns. In addition to replicating the specific increase in continuous ratings within categories, we note that the size of the average percent change varies between categories (Fig. 5a). The largest increase in ratings is observed for fractal image triplets. These increased momentary ratings may be related to the degree of surprise (Alba & Williams, 2013; Schultz, Dayan, & Montague, 1997) associated with viewing unexpected (and pleasant) changes in imagery. Indeed, our fractal images are arguably more dissimilar within their category than our human poses are (Appendix Fig. 1). Furthermore, expectations for specific movements formed while viewing static body postures (Freyd, 1983) may dampen the perceived novelty of subsequent changes in dance imagery. Future work might extend the methods proposed here across different domains and timescales. This would allow us to further understand the dynamics of aesthetic response within different stimulus categories.

The choice preference for ABA sequences is also confirmed for both dance pose categories. However, for fractal patterns, participants indicate no significant preference when choosing between ABA and AAA sequences (Fig. 5b). Here a natural explanation is that an additional preference for symmetry is evoked for abstract and geometric patterns (Cárdenas & Harris, 2006; Makin, Pechcinenda, & Bertamini, 2012; Tyler, 1995). For instance, consider an additional preference for reflexive symmetry (at the midline) across a fractal triplet (recall that these choices are made between sequences displayed as panels, as in Fig. 1b). This sequence-wise symmetry would be disrupted by introducing a middle element change (as in the case of ABA arrangements) featuring some local (within-image) asymmetry. Indeed, 7 out of 12 of our individual fractals happen to be asymmetric in terms of their displayed shapes (Appendix Fig. 1). Under this interpretation, ABA fractal sequences should be preferred over AAA sequences, after controlling for local asymmetries.

We note that we are unable to detect any additional linear sequence effects when performing category-wise regression analyses on individuals’ discrete rating trials (Appendix Fig. 4). The absence of such effects may be attributed to a lack of statistical power, though we note again that contrast effects arising from the previous trial are found in various modalities (Cogan et al., 2013; Pegors et al., 2015; Toussignant & Bodner, 2014) as well as in our main analysis.

We conclude that aesthetic responses toward arrangements might offer a normative explanation as to why particular creative norms persist. For instance, song structures may have been preserved over time because of their exploitation of particularly favored sequences. On this basis, it seems likely that the psychological mechanisms described here also shape the presentation formats of widely appreciated art. Future theories of aesthetic value should accommodate these factors, in addition to stimulus-specific responses involved in aesthetic experience (e.g., emotional and embodied responses).

Appendix A. Supplementary Files and Figures

Supplementary figures, stimulus materials, and results to this article can be found online at https://doi.org/10.1016/j.actpsy.2018.04.018.

References


appreciation. Frontiers in Human Neuroscience, 10.