[This manuscript was accepted for publication by Psychology of *Aesthetics, Creativity, and the Arts* on 26 May, 2022]

# The Aesthetic Quality Model:

# Complexity and Randomness as Foundations of Visual Beauty by Signaling Quality

Dario Krpan[1] and Wijnand A. P. van Tilburg[2]

[1] Department of Psychological and Behavioural Science, London School of Economics and Political Science

[2] Department of Psychology, University of Essex

## Author Note

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all participants.

The authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. Authors contributed equally to this work.

Correspondence concerning this article should be addressed to Dario Krpan; Department of Psychological and Behavioural Science; London School of Economics and Political Science; London, WC2A 2AE; United Kingdom; Email: d.krpan@lse.ac.uk

## Abstract

Visual complexity has been identified as a fundamental property that shapes the beauty of visual images. However, its exact influence on beauty judgments, and the mechanism behind this influence, remain a conundrum. In the present article, we developed and empirically evaluated the *Aesthetic Quality Model*, which proposes that the link between complexity and beauty depends on another key visual property—randomness. According to the model, beauty judgements are determined by an interaction between these two properties, with more beautiful patterns featuring comparatively high complexity and low randomness. The model further posits that this configuration of complexity and randomness leads to higher beauty because it signals quality (i.e., creativity and skill). Study 1 confirmed that black and white binary patterns were judged as more beautiful when they combined high complexity with low randomness. Study 2 replicated these findings using an experimental method and with a more representative set of patterns, and it pointed to quality attribution as a candidate mechanism underlying the beauty judgements. Studies 3 and 4 confirmed these findings using experimental manipulation of the mechanism. Overall, the present research supports the aesthetic quality model, breaking new ground in understanding the fundamentals of beauty judgement.

*Keywords:* Beauty, computational aesthetics, complexity, randomness

# The Aesthetic Quality Model:

## Complexity and Randomness as Foundations of Visual Beauty by Signaling Quality

Some of the core characteristics of art are its diversity and emphasis on uniqueness of styles and composition. However, underneath this subjective façade, are there underlying properties that can explain when and why some works of art are perceived as more beautiful than others? For millennia, this question has captivated intellectuals across domains, ranging from philosophy and science to psychology and art. For example, during the Renaissance, creators such as Leonardo da Vinci believed that the proportions described by the golden ratio comprise the essence of beauty (Di Dio, Macaluso, & Rizzolatti, 2007). Johann Sebastian Bach embedded within his brilliant compositions symmetrical devices whose intricacy has long fascinated scholars (e.g., Jander, 1991; Hofstadter, 1979). Kant, on the other hand, posited that beauty does not reside in the work of art itself but in the interplay of the observer's imagination and subjective associations inspired by the artwork (Kant, 2000, Wicks, 1995).

Psychologists have studied how inferred or observed attributes of artists serve as basis for beauty judgements, such as their perceived creativity (Hager, Hagemann, Danner, & Schankin, 2012) or skill (Van Tilburg & Igou, 2014). Others have examined how cognitive processes, such as processing fluency (Schwarz & Winkielman, 2004), contribute to aesthetic pleasure. However, less is known about how objective features of the judged work shape beauty judgements, perhaps mediated by the psychological processes just mentioned. Several overarching models that have been proposed in this respect typically identify image complexity as one of the key visual features (e.g., Leder, Belke, Oeberst, & Augustin, 2004; Leder & Nadal, 2014; Palmer, Schloss, & Sammartino, 2013; Pelowski, Markey, Forster, Gerger, & Leder, 2017). However, whereas the models generally agree that complexity is important for aesthetic appreciation, the exact link between this visual element and beauty judgments remains a conundrum. For example, the core hypothesis in this regard—that complexity and beauty have an inverted-U relationship, and most aesthetically pleasing images are thus the ones with medium levels of complexity (e.g., Berlyne, 1970)—has been extensively tested but produced mixed results (e.g., Silvia, 2005; Van Geert & Wagemans, 2020). Overall, to our knowledge, an empirically supported model that resolves the link between complexity and aesthetic appreciation and outlines a clear mechanism behind this link has not yet been developed (see Van Geert & Wagemans, 2020).

In the present article, we propose a foundational model of beauty judgements called the *Aesthetic Quality Model*, which indicates that understanding the link between complexity and beauty requires introducing randomness into the equation. The model postulates that beauty judgements are, in part, a function of these two visual qualities—most beautiful images are the ones that are both complex and characterized by order. Furthermore, this model proposes that quality attributions—the impression that something requires skill and creativity to (re)recreate—mediate this impact of the complexity by randomness contingency onto beauty judgements.

After introducing our model, we test it in the context of beauty judgements for black and white binary patterns. These simple stimuli are employed for two reasons. First, they allow precise computation and manipulation of complexity and randomness. Second, for these patterns, it is possible to generate a representative set of the stimuli that can occur in "nature" and therefore ascertain generalizability of the findings. This resolves one of the problems encountered in previous research, where stimuli typically involved a selection of artworks or patterns for which it was not determined whether they constitute a generalizable sample of all such stimuli or merely their rare manifestations. Indeed, relying on a set of stimuli that are not a representative sample can severely bias the findings and lead to false conclusions about the existence of a phenomenon (Westfall, Judd, & Kenny, 2015).

## Complexity and Aesthetic Appreciation

Visual complexity has been operationalized in many ways depending on the types of images studied (e.g., Chipman, 1977; Donderi, 2006; Jakesch & Leder, 2015; Sherman, Grabowecky, & Suzuki, 2015; Van Geert & Wagemans, 2020). A broad definition that contains the essence of these different operationalizations is that complexity comprises "the quantity and variety of information in a stimulus" (Van Geert & Wagemans, 2020, p. 135). This visual quality can therefore be captured via a variety of measures, from the information theoretic ones, such as the number of bits needed to encode an image (e.g., Mather, 2018), to the ones that quantify observable image characteristics, such as the number of individual elements that constitute it (e.g., Tinio & Leder, 2009). It is important to emphasize that there does not seem to be one core measure of complexity, given that different measures can capture its core features in different ways and may be suitable for different image types. Therefore, the broad definition that captures variety and quantity of elements within an image is necessary to account for the variety of measurements available by summarizing the essence they all share. When this definition is applied to the type of binary patterns used in the present research (Figure 1), complexity can be described using an overarching operationalization that is intuitive and yet inclusive of various measures that are strongly predictive of how people subjectively perceive this visual quality: as the amount of different constituent components that can be recognized in a pattern (Chipman, 1977). For example, as can be seen in Figure 1, each of the two high-complexity patterns has a larger number of separate black shapes (i.e., areas consisting of one black square or several black squares where there is no visible vertical or horizontal border between them) than the two low-complexity patterns, even if the total area the shapes cover is the same in each pattern.

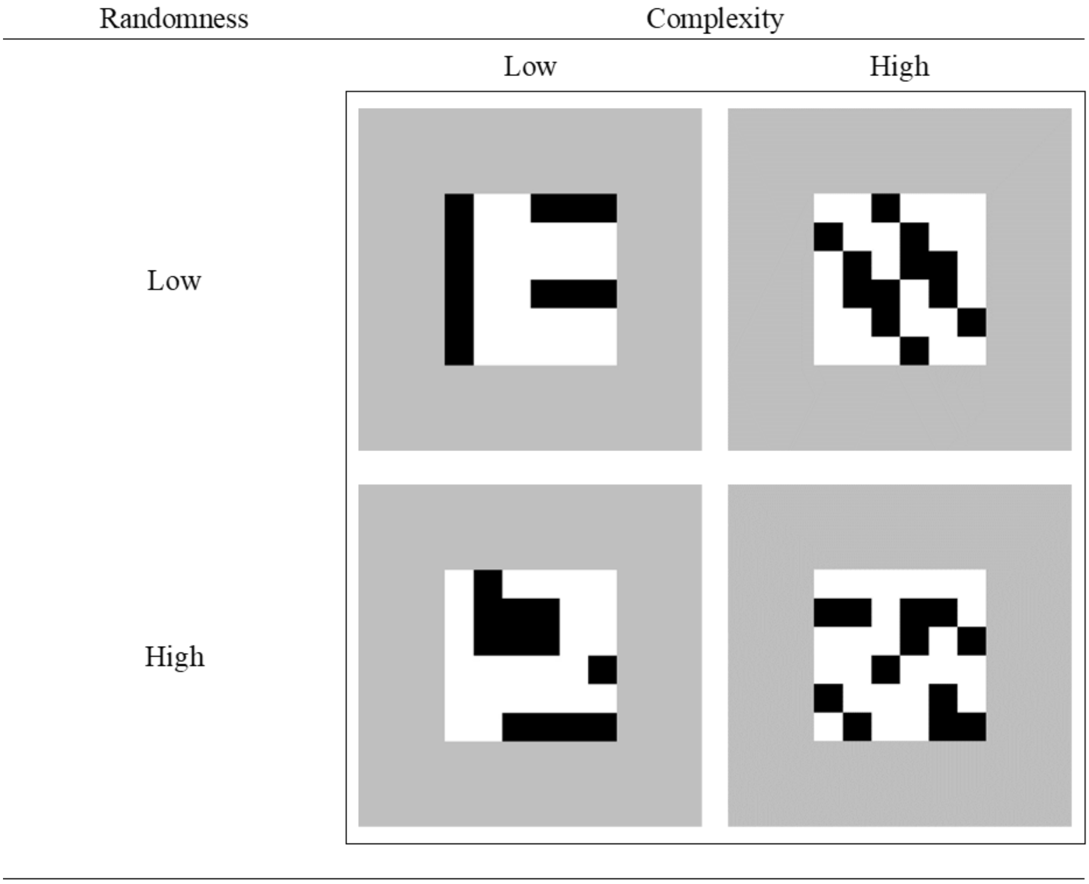


*Figure 1*. Examples of Complex and Random Patterns.

Berlyne (1963, 1970, 1973, 1974; see also Birkhoff, 1932) found inspiration in Wundt's (1874) inverted-U curve and the Yerkes-Dodson law (Teigen, 1994) to propose that moderately complex stimuli are perceived as most beautiful because they lead to moderate levels of arousal that are maximally rewarding. Some studies indeed found an inverted-U relationship between complexity and beauty (Forsythe, Nadal, Sheehy, Cela-Conde, & Sawey, 2011; Hekkert & Van Wieringen, 1990; Imamoglu, Ç., 2000; Güçlütürk, Jacobs, & van Lier, 2016). However, findings have been mixed (Silvia, 2005; Van Geert & Wagemans, 2021). Whereas some studies found no relationship between complexity and beauty (Nadal, Munar, Marty, & Cela-Conde, 2010), other studies found that more complex stimuli are more aesthetically pleasing (Friedenberg & Liby, 2016; Mayer & Landwehr, 2018). Therefore, research overall indicates that complexity considered in isolation may be insufficient as basis for understanding beauty judgements.

## Introducing Randomness

Several researchers have suggested that randomness, another visual quality that has been studied in relation to aesthetic preferences, may need to be considered to achieve a more nuanced understanding of the link between complexity and beauty judgments (Chipman, 2013; Van Geert & Wagemans, 2020, 2021). In the literature, the terms randomness, disorder, or disorganization are used relatively interchangeably and broadly capture the extent to which visual elements of an image lack some underlying order or organization; appearing to be randomly scattered (Falk & Konold, 1997; Van Geert & Wagemans, 2020, 2021). As with complexity, there does not seem to be one core measure of randomness (Van Geert & Wagemans, 2020). Indeed, various measures, from how different the number of black pixels is across different subsections of a binarized image (Hübner and Fillinger, 2016), to how frequently basic building blocks of an image alter (Falk & Konold, 1997), can capture its essence and be more suitable for different image types. In the context of the type of binary patterns on which we focus, randomness can be described as the extent to which the arrangement of the black squares lacks some easily identifiable underlying principle.[1] As a simple intuitive illustration (see Figure 1), if we count the number of black squares in each row of the two low randomness patterns (starting with the top row), we will see that their quantities regularly repeat (for the first low randomness pattern, these quantities are 4-1-1-4-1-1, and for the second one they are 1-2-3-3-2-1). In contrast, for the two high randomness patterns, the quantities are 1-3-3-1-0-4 and 0-4-2-1-2-3 respectively, and it is difficult to

[1] One visual characteristic that is sometimes mentioned alongside randomness is symmetry, given that it is considered an aspect of order that can signal low randomness (Van Geert & Wagemans, 2020). However, when defining randomness, we do not refer to symmetry because this characteristic was also established as one of the strongest predictors of perceived complexity (Chipman, 1977) and therefore cannot be used to make a clear distinction between complexity and randomness.

identify some underlying principle behind their variation. A visual consequence of this is that the patterns seem disorganized.

Given that complexity and randomness tend to be conflated frequently in the literature, and that random images may to some appear as more complex than non-random ones (Chipman, 1977; Donderi, 2006; Falk & Konold, 1997), it is important to further clarify the distinction between complexity and randomness, and to integrate the two constructs. As illustrated in Figure 1, whereas patterns that belong to the same complexity category tend to have a similar number of discernible visual elements (i.e., areas consisting of an individual black square or several black squares where there is no visible vertical or horizontal border between them), in highly random patterns these elements appear disorganized and arbitrarily constructed, without an easily identifiable underlying principle, whereas in low randomness patterns the elements are arranged in a seemingly orderly manner. Therefore, a relatively complex pattern can be both random or non-random, and the same logic applies to less complex patterns. Here, the simple intuitive example described when introducing randomness that involves counting the number of black squares in each row can again be evoked to illustrate how it is possible to distinguish between the complex low versus high randomness patterns, or between the non-complex low versus high randomness patterns.

So far, few studies have investigated how different combinations of randomness and complexity shape aesthetic perception (Westphal-Fitch & Fitch, 2017). For example, Van Geert and Wagemans (2021) studied the perception of colored compositions. However, they focused on how soothing and fascinating people found these compositions, rather than on beauty judgments. Moreover, whereas they demonstrated that complexity and randomness independently predicted these dependent variables, their research did not generate findings that would explain a joint role of the two visual qualities in aesthetic perception. Overall, although it has been speculated that complexity and randomness may hold a key to understanding beauty (Gabriel & Quillien, 2019; Van Geert & Wagemans, 2020), their exact relationship to aesthetic preferences and the mechanism behind this relationship remain unclear. In the next section, we propose an aesthetic quality model that may explain how and why the two qualities interact in shaping beauty.

## Aesthetic Quality Model

The model we propose rests on the basic assumption that people's aesthetic preferences (e.g., beauty) are grounded in their perception of artistic quality of a work of art (e.g., Hagtvedt, Patrick, & Hagtvedt, 2008; Kozbelt, 2004; Van Tilburg & Igou, 2014). Although there are many characteristics of an artwork that can determine quality, two key components that have been established by previous research are creativity and skill (Kozbelt, 2004). Art is generally perceived as a creative endeavor, and skill is seen as a prerequisite for creating an image that is original and unique (Newman & Bloom, 2012). For that reason, perceived skill and creativity of visual images are typically highly correlated (Chan & Zhao, 2010; Hekkert & Van Wieringen, 1996) and comprise an underlying construct that has been labeled aesthetic quality (Kozbelt, 2004; see also Christensen, Ball, & Reber, 2020). Several studies indicate that perceived quality covaries with aesthetic appreciation (Hagtvedt et al., 2008). For example, the subjective quality of images was correlated with liking for both expert and non-expert raters (Pelowski, Markey, Goller, Förster, & Leder, 2018). Moreover, in a scale measuring aesthetic perception, quality was the factor that had strongest relationship with positive attraction (i.e., beauty; Hager et al., 2012).

Our aesthetic quality model treats perceived quality as a mediator in the link from randomness and complexity, on the one hand, to aesthetic judgement, on the other. Specifically, we posit that people interpret combinations of complexity and randomness in terms of quality, which then shapes beauty judgments.

To start with complexity, how and why might this attribute shape quality perceptions? Simple (i.e., non-complex) images may indicate a lack of perceived quality based on the impression that it might be easy to produce them, thus requiring little creativity and skill (e.g., Kruger, Wirtz, Van Boven, & Altermatt, 2004; Newman & Bloom, 2012). Indeed, previous research supports this assumption—even if testing the link between complexity and beauty has produced mixed findings for medium and high levels of complexity, simple patterns are consistently perceived as least appealing (e.g., Forsythe et al., 2011; Friedenberg & Liby, 2016; Mayer & Landwehr, 2018).[2]

While the above offers a tentative account for the perceived beauty of patterns low in complexity, how might patterns of higher levels of complexity relate to perceived quality and aesthetic judgment, especially considering prior mixed findings? We propose that a key factor is the patterns' randomness. Disordered images are generally perceived as less beautiful than the ordered ones (e.g., Bertamini, Makin, & Rampone, 2013; Makin, Pecchinenda, & Bertamini, 2012; Westphal-Fitch & Fitch, 2017). Within our aesthetic quality model, we posit that this is the case because disorder is associated with chance and the absence of creative process that requires skill (Falk & Konold, 1997; Gabriel & Quillien, 2019; Serafin, Kozbelt, Seidel, & Dolese, 2011). While it is therefore possible that non-random (vs. random) simple patterns may be perceived as more beautiful in some instances, this difference may either be small or nonexistent, given that, as we have argued, simple patterns may generally indicate a lack of perceived quality (e.g., Kruger et al., 2004). For more complex patterns, however, randomness should play a vital role in determining beauty. Disordered complex patterns may be judged as less beautiful because of the impression that

[2] It is important to emphasize that for every research finding on the link between a visual quality and beauty there are likely exceptions to the rule. For example, certain art forms such as haiku may be specifically valued based on their simplicity despite the general finding that simple patterns tend to be least appealing. Overall, our predictions concerning the aesthetic quality model describe how beauty perception functions on average but do not imply that there are no exceptions in this regard.

their creation does not require levels of creativity and skill that would be indicative of high quality (Gabriel & Quillien, 2019; Newman & Bloom, 2012; Serafin et al., 2011). In contrast, complex non-random patterns should be linked to beauty because turning random complexity into order may require a creative and skillful effort that characterizes creation.

A notion similar to the above has, for example, been explored in architecture, where the construct of "well-ordered complexity" has been proposed to explain the beauty of buildings and city designs (Gabriel & Quillien, 2019). Although we have not identified any published empirical findings in the literature that would directly support our prediction, an unpublished study by Chipman (2013) is consistent with our model. She showed that, for patterns that were classified as structured (i.e., low in randomness) in her previous research (Chipman, 1977), the positive relationship between complexity and aesthetic quality was stronger than for the unstructured patterns. However, one of the limitations in this regard is that the unstructured (vs. structured) patterns used as stimuli on average had considerably higher levels of complexity, and our assumption that highly complex but structured binary patterns would be perceived as the most beautiful ones therefore remains untested.

Overall, based on the present theorizing, our model combines two key arguments. First, complexity and randomness should interact in predicting beauty. That is, complex non-random patterns should be perceived as the most beautiful ones (i.e., more beautiful than complex random patterns or either random or non-random patterns of low complexity). Second, the attribution of quality (i.e., creativity and skill) is a key psychological mediator and transfers the influence of the interaction between complexity and randomness on perceived beauty. Note that beauty, complexity, and randomness are in this context necessarily comparative judgements relative to other members of a defined population of patterns, given that aesthetic judgment typically does not happen in isolation and depends on other images that serve as reference points (e.g., Chipman, 1977; Forsythe et al., 2011; Van Geert & Wagemans, 2020, 2021).

## Overview of the Present Research

Our aesthetic quality model posits that aesthetics judgements rest for an important part on the attribution of perceived 'quality'—skill and creativity—to visual patterns. Such attributions are in turn based on the relative complexity and randomness that patterns feature. Low randomness, or disorder, and high complexity reflect that patterns are in essence rare: their occurrence seems hardly due to chance but instead suggests the outcome of a required skillful and creative process. In all, we propose that aesthetically pleasing visual patterns tend to be characterized by relatively high complexity and low randomness.

We derive four hypotheses from our model: Combining low randomness with high complexity produces visual patterns that are comparatively aesthetically pleasing, corresponding to a randomness × complexity interaction on beauty judgements (Hypothesis 1). Furthermore, we propose that attributions of perceived quality—finding patterns skillful and creative—act as mediators: visual patterns characterized by low randomness and high complexity compel viewers to attribute high quality to them, representing a randomness × complexity interaction on perceived quality (Hypothesis 2). This attribution of quality in turn results in corresponding positive aesthetic judgments; a positive association between attributed quality and aesthetic judgement (Hypothesis 3), cumulating in a pattern of mediated moderation where the interactive impact of randomness and complexity in aesthetic judgements is 'transmitted' by quality attributions (Hypothesis 4).

We evaluated our hypotheses in a series of four empirical studies. Specifically, in Study 1 we tested, using a correlational design, if the aesthetic judgement of visual patterns was a function of an interaction between randomness and complexity (Hypothesis 1), and did so using a range of objective and subjective indicators of these two factors. Study 2 also tested for the existence of this interaction (Hypothesis 1) but did so in an experimental design where we orthogonally manipulated randomness and complexity and relied on an improved set of visual patterns. Furthermore, we measured quality attributions and tested if these varied as a function of the same randomness × complexity interaction (Hypothesis 2), if these quality attributions were positively associated with aesthetic judgements (Hypothesis 3), and if quality attributions statistically mediated the interaction effect on aesthetic judgements (Hypothesis 4). In the ensuing Studies 3 and 4, rather than testing Hypotheses 2-4 using a mediation approach, we manipulated the alleged psychological process, attributed quality, directly to gauge its causal role in the proposed mechanism (Spencer, Zanna, & Fong, 2005).

## Study 1

We first tested the hypothesis that complexity and randomness interact in predicting beauty (i.e., most beautiful patterns should be the ones that have low randomness and high complexity) on a set of 45 patterns adopted from Chipman (1977). These stimuli were selected because they contained a range of patterns of varying complexity levels that we found optimal for preliminary tests of our core hypotheses. More precisely, these stimuli were divided into 15 complex patterns, 15 simple patterns, and 15 “basic” patterns that comprised a range of complexity levels from low to high. All participants rated complexity, randomness, and beauty for all 45 patterns, and the hypothesis was then tested using multilevel models (Finch, Bolin, & Kelley, 2019; Hayes, 2006). In addition to probing Hypothesis 1 on participants’ *subjective* complexity and randomness ratings, we tested it using two *objective* indicators of these visual qualities.

In contrast to the original research by Chipman (1977), all the patterns in our study were presented to participants digitally, on the computer screens, rather than in a printed version. We used this method because digitally presenting information has become ubiquitous in the current digital age, and also because Chipman (1977) showed that either the size of the patterns or the context in which they were presented had no influence on participants' complexity ratings. To verify that the different mode of presentation and stimuli sizes in our study indeed did not confound pattern perception, we obtained participants' complexity ratings from the original research by Chipman (1977) to compare them with the ratings from the present study (note that Chipman did not assess perceived randomness).

Moreover, to ensure that participants' perception of pattern qualities did not depend on a specific rating procedure, we used two different procedures in the present research to probe whether they generate different results. Half of the sample used a scoring method in which participants assigned different relative numbers to patterns (Chipman, 1977) to express how they perceived them concerning a specific quality (e.g., complexity, randomness, or beauty). The other half simply rated each quality using a slider (with values ranging from 0 to 100).

In addition to the three qualities important for hypothesis testing (i.e., complexity, beauty, and randomness), we also asked participants to rate additional qualities (i.e., boredom, positivity, negativity, busyness, and intensity) to decrease the likelihood that they understand the specific predictions we were testing, thus reducing potential experimenter demand effects (Orne, 1962, 2009), but also to inform our other research. Finally, we measured several exploratory personality variables to probe whether personality shapes the interaction between complexity and randomness in predicting beauty. More specifically, we measured the BIG5 personality traits (Friedenberg, 2019; Gosling, Rentfrow, & Swann Jr, 2003), political orientation (liberal versus conservative; Graham, Haidt, & Nosek, 2009), boredom proneness (Struk, Carriere, Cheyne, Danckert, 2017), open mindedness (Haran, Ritov, & Mellers, 2013), and need for closure (Roets & Van Hiel, 2011), because previous research indicated that these individual differences may be linked to aesthetic preferences (e.g., Chirumbolo, Brizi, Mastandrea, & Mannetti, 2014; Furnham & Rao, 2002; Furnham & Walker, 2001a, 2001b; Kandler et al., 2016; Mastandrea, Bartoli, & Bove, 2009; Ostrofsky &

Shobe, 2015; Rawlings, 2000; Rosenbloom, 2006; Swami & Furnham, 2012; Wiersema, Van Der Schalk, & van Kleef, 2012; Wilson, Ausman, & Mathews, 1973).

## Method

### *Transparency and Openness*

For all studies in this article, we report all data exclusions, all manipulations, and all measures. For each study, the section *Determining Sample Size* outlines the rationale behind the sample size. All data, analysis codes, and research materials are available via the Open Science Framework (OSF), using the following link: https://osf.io/n7p5z/. None of the studies in the present article were pre-registered.

### *Stimuli*

The stimuli in the present study were 45 black and white patterns from Chipman (1977; Experiment 1). All patterns consisted of 6 × 6 squares, 12 of which were black and 24 white. The size of all patterns was 499 (width) × 499 (height) pixels, and they were presented on a gray surface sized 997 (width) × 997 (height) pixels (see Figure 2 for an example).

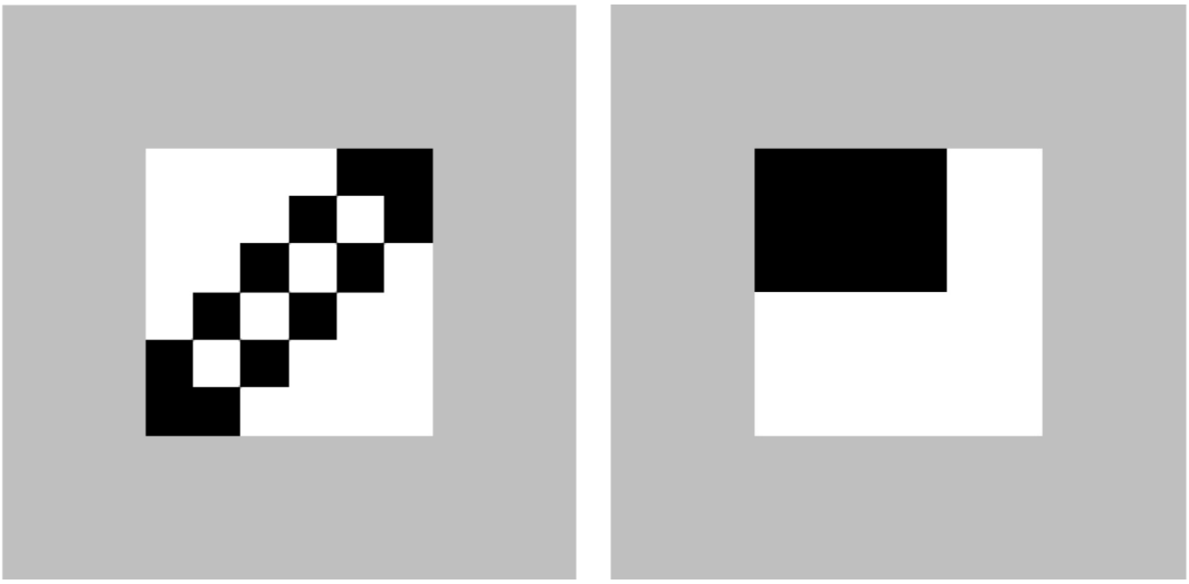


*Figure 2.* Examples of patterns that were used as stimuli in Study 1. The pattern on the left was rated as the most beautiful by participants, and the pattern on the right as the least beautiful.

### *Determining Sample Size*

Given that the statistical power of multilevel models is shaped by a range of different factors, some of which are still being investigated, and that many of the parameters required cannot be reliably determined in advance (Mathieu, Aguinis, Culpepper, & Chen, 2012), we relied on simulations reported by other researchers to determine adequate sample size for the present study. Research generally agrees that level-1 variable sample size (i.e., in our case, the number of patterns each participant rated: 45) is more important than level-2 variable sample size (i.e., in our case, the number of participants tested) for determining power, given that high power cannot be achieved even with a large sample size for level-2 variables if the sample size for level-1 variables is small (Aguinis, Gottfredson, & Culpepper, 2013; Lane & Hennes, 2018; Maas & Hox, 2005; Mathieu et al., 2012; Scherbaum & Ferreter, 2009). Mathieu et al. (2012) showed that, for effects that are typically obtained in the literature, when 18 observations are measured per level-1 variable, a sample size of 60 for level-2 variable leads to a large power (1 – β >.95), even for cross-level interactions that are typically more demanding. Similarly, Maas and Hox (2005) showed that level-2 sample sizes of 50 participants or less may lead to biased estimates. Based on these findings, and given that the number of observations per our level-1 variable was relatively large (i.e., 45), we concluded that recruiting 60 or more participants in total would be sufficient for testing the hypothesis. To be on the safe side, we decided to recruit a sample that was roughly three times larger (i.e., between 180-200 participants). The final sample size obtained was 193 participants (see the *Participants and Design* section below).

### ***Participants and Design***

One hundred and ninety-three participants completed the study (Female = 82, Male = 111; $M_{\text{age}}$ = 37.641; $SD_{\text{age}}$ = 11.572) via Amazon Mechanical Turk (MTurk). They identified their nationalities as American, Asian, Canadian, Filipino, Guyanese, Hispanic, Indian, Polish-American, and Slovak. Payment was $3.00. To ensure high-quality responses, we recruited only the workers who were awarded the "Masters" qualification on MTurk based on various quality indicators (e.g., approval rates). All participants viewed the 45 patterns and rated them on the following qualities: complexity, beauty, boredom, positivity, negativity, busyness, intensity, and randomness. Participants were randomly assigned into two different quality rating procedures: number versus slider (see the *Procedure* section below). Seventy-eight participants eventually completed the study in the number rating procedure and 115 in the slider rating procedure.  This and other studies in the article were approved by the Research Ethics Committee of the university of one of the authors.

### ***Procedure***

The study was administered via Qualtrics. After giving consent, participants reported demographics (see the *Measures* section below) and subsequently received detailed instructions. They were told that they would be asked to score 45 black and white patterns on complexity, beauty, boredom, positivity, negativity, busyness, intensity, and randomness. Complexity referred to how complex a pattern seemed to them; beauty to how beautiful they found it; boredom to how boring they found it; positivity to whether the pattern made them experience positive feelings; negativity to whether it made them experience negative feelings; busyness to how visually busy the pattern seemed; intensity to whether the pattern produced intense sensations while they were looking at it; and randomness to how random they found it (i.e., to what extent it lacked any underlying order).

Participants were randomly allocated into two different pattern rating procedures. In the number rating condition (see Chipman, 1977), they were told to give the first pattern whatever number that corresponds to how they perceive it in terms of each of the qualities. Then, they were instructed to give the next pattern a number that corresponds to how they perceive it regarding a quality in relation to the previous patterns (e.g., “If the next pattern seems twice as complex, give it a complexity number twice as large. Alternatively, if it is half as complex, give it a complexity number half as large.”). Participants were told to use whatever numbers are necessary to represent the relationship between the patterns, and it was emphasized that it is important to understand that they should use a previous pattern as the reference when scoring the next pattern. In the slider rating condition, participants were told that, for each quality, a slider with values ranging from 0 to 100 would be displayed, and they would need to adjust the sliders to correspond to how they you perceive a pattern regarding each of these qualities. It was explained that a score of 0 (100) corresponds to the pattern being very low (high) on a specific quality. Participants were told to use any number ranging from 0 to 100 that corresponds to how they perceive it in terms of a specific quality. Importantly, participants in either condition were told to first look at all the patterns to get a general idea about their appearance and only then start rating each pattern.

After rating the patterns, participants filled in various exploratory individual-differences measures (see the *Measures* section below). Finally, they were debriefed and received a seriousness check (Aust, Diedenhofen, Ullrich, & Musch, 2013).

### *Measures*

**Dependent Variable.** The dependent variable was *perceived beauty*. To compute this variable, we first transformed a participant's beauty ratings of all 45 patterns into ranks. This procedure was used because previous studies with similar design used a comparable scoring (e.g., Chipman, 1977) and because it allowed us to analyze beauty ratings for participants from different rating conditions (number vs. slider) together and to compare them. An average rank was assigned to duplicate scores. Higher ranks indicated higher beauty.

**Predictor Variables: Subjective Complexity and Randomness.** These variables were computed using the same procedure as *perceived beauty*. Participants' raw complexity and randomness scores were transformed into ranks, and duplicate scores were assigned an average rank. Higher ranks indicate higher complexity and randomness.

**Predictor Variables: Objective Complexity and Randomness.** To assess objective complexity and randomness, for each of the 45 stimuli patterns we first computed the most robust measures of these qualities identified by previous research. Given that many complexity and randomness measures have been proposed, we focused on those developed specifically for two-dimensional black and white patterns comparable to the ones used in the present research. Next, we analyzed which of these measures were the best predictors of participants' subjective complexity and randomness ratings: one strongest predictor of subjective complexity and one strongest predictor of subjective randomness were therefore selected as the best objective measures of these qualities to be used in testing Hypothesis 1. The computed measures and the validation procedure are extensively explained in Supplementary Materials (pp.3-9); below we present the two final predictors we selected.

As an indicator of objective complexity, we selected *turns* (Chipman, 1977), which captures complexity by identifying the number of corners present in a pattern consisting of black and white squares. Namely, if two neighboring sides of a black square are on the boundary between black and white, one turn is counted. Higher number of turns indicates larger complexity.

As an indicator of objective randomness, we used a measure that we developed based on Fourier transformations and therefore labelled it *Fourier randomness*. This measure essentially reflects whether an image contains a small set of comparatively pronounced square waves (indicative of low randomness). Discrete Fourier transform (e.g., Winograd, 1978) breaks down functions with a finite range—such as a complex wave or pattern—into a series of basic sinusoids, illustrated in Figure 3. This transformation can be applied to two-dimensional functions or functions of higher dimensional order, such as the three-dimensional 6 × 6 patterns we used. The result of this decomposition in the context of our 6 × 6 patterns is a series of 36 square waves, each characterized by a vertical and horizontal frequency and an amplitude expressed as a complex number. This is illustrated in Figure 4 for one of the presumably 'nonrandom' Chipman patterns and in Figure 5 for one of the possibly more 'random' patterns.

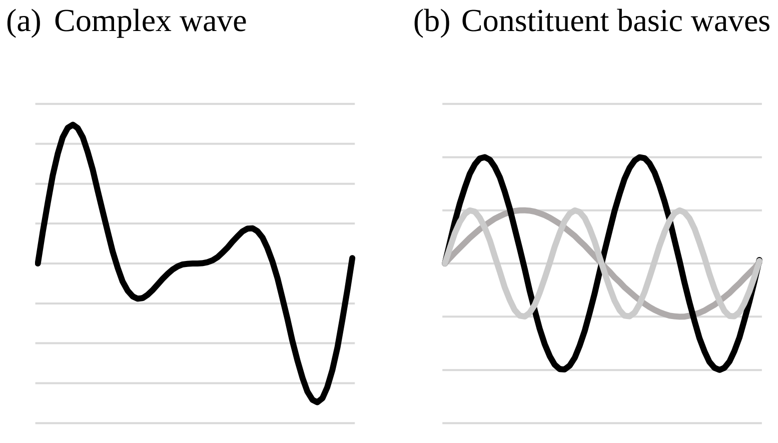


*Figure 3*. Decomposition of a complex wave into basic waves

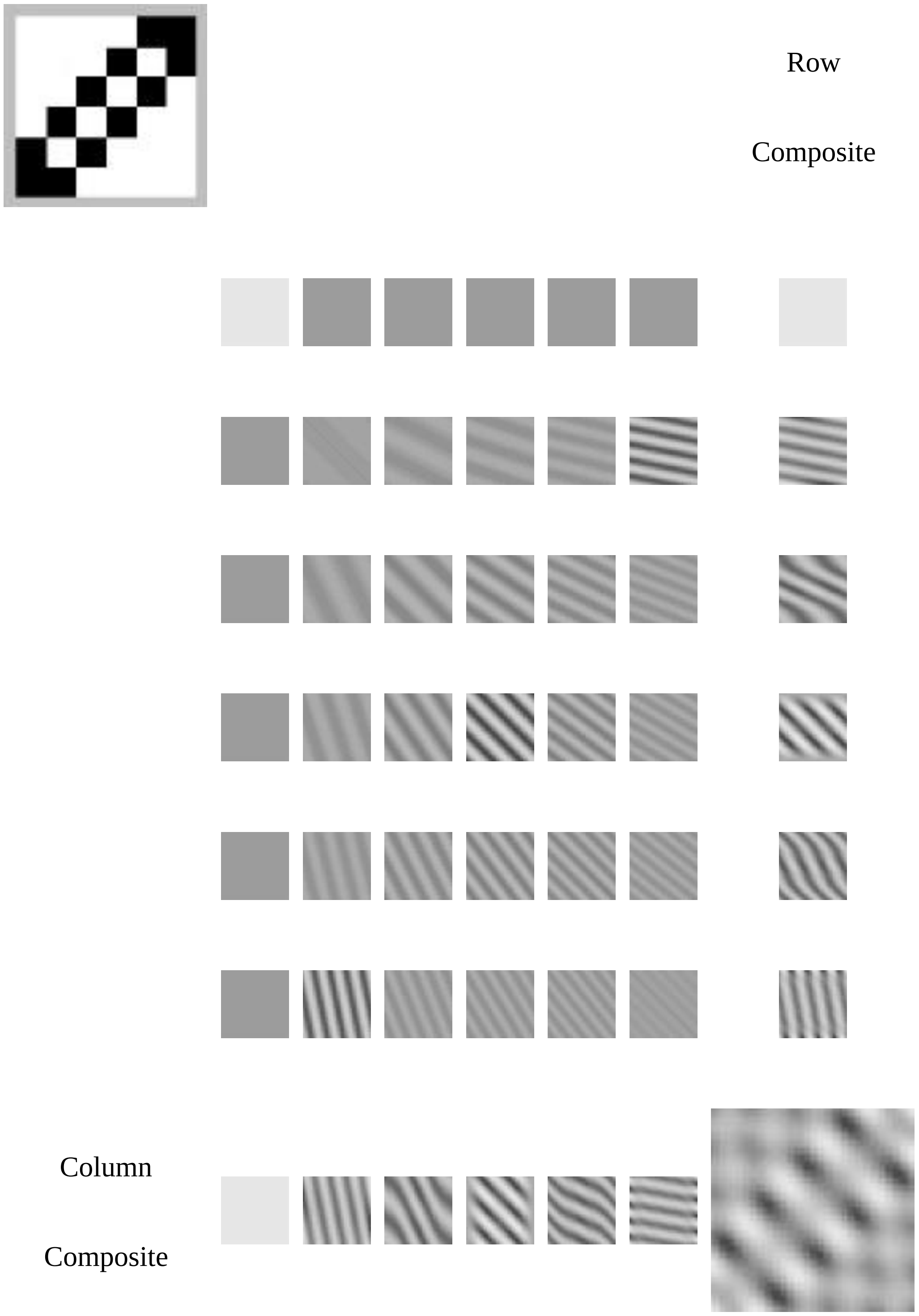


*Figure 4.* Discrete Fourier transform of a 6 × 6 pattern presumably low on randomness. Values in the reproduced image match those of the original image at corresponding coordinate intersections.

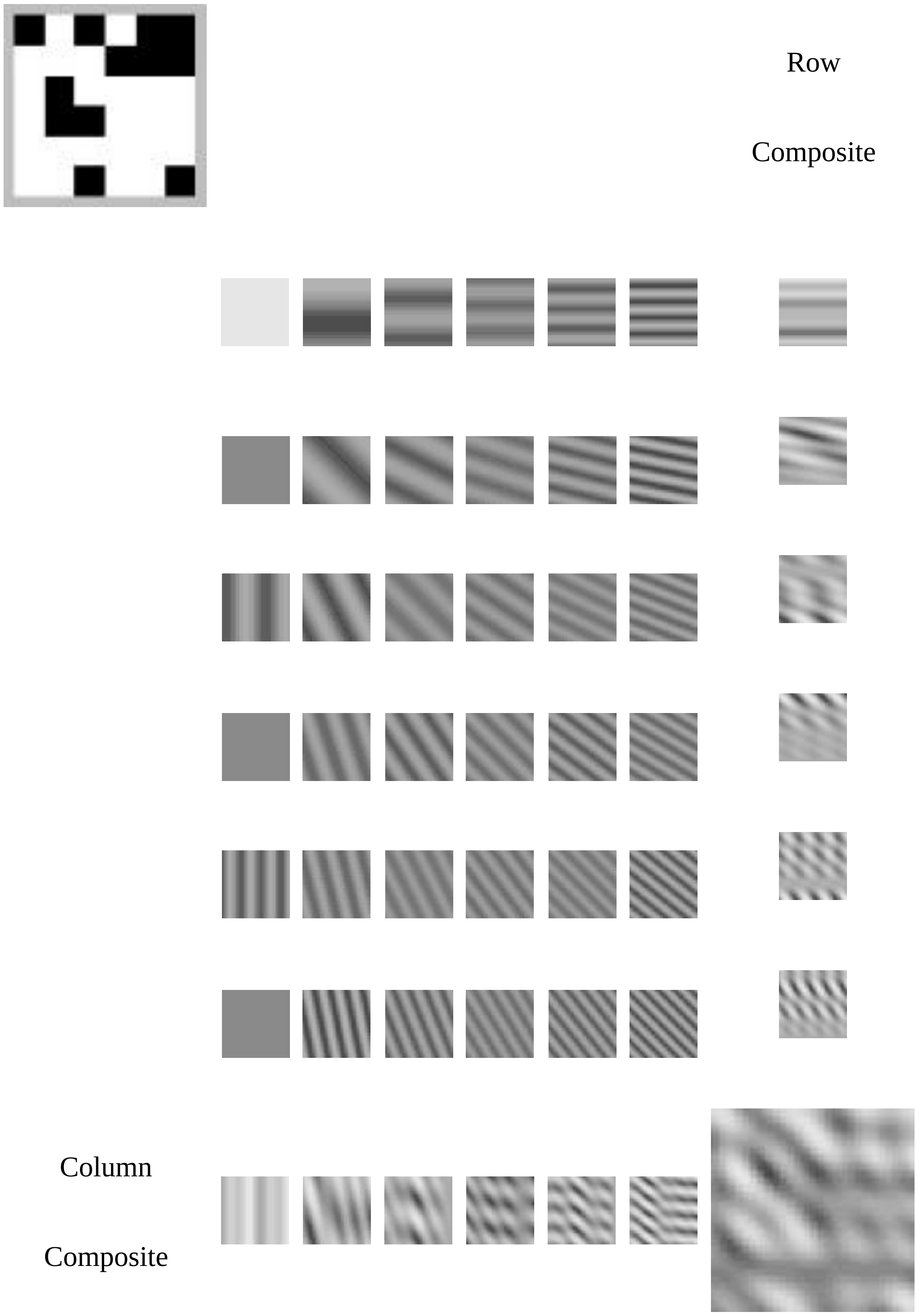


*Figure 5.* Discrete Fourier transform of a 6 × 6 pattern presumably high on randomness. Values in the reproduced image match those of the original image at corresponding coordinate intersections.

Retaining all its resultant waves, a Fourier transform reproduces the original image perfectly. However, not all waves contribute to this to the same degree. That is, some waves exert a stronger influence on this reproduction than others, evident from their comparatively high intensities. This feature of Fourier transform, that the intensity of some waves is larger or smaller than others, can be used as a means for filtering out noise (e.g., in audio or image processing; Kutay, & Ozaktas, 1998; Tempelaars, 1996). As an example, the prototypical disorderly state of 'white noise' is characterized by waves of equal intensity across frequencies. Non-random patterns, on the other hand, tend to feature waves of different intensities, with those of high intensity being particularly characteristic of a strong reoccurring pattern. This feature might be helpfully utilized to quantify how disordered a pattern is, for example by calculating how many waves with comparatively large intensity emerge; the lower this number, the less noisy, or random, a pattern appears to be. We treated the intensity of a given wave as comparatively 'large' if the magnitude of its amplitude, calculated using the conjugate given that these are complex numbers, amounted to over 10% of the sum of all these amplitudes (Mayer, Khairy, & Howard, 2010). We excluded from this calculation the first square wave given that this characterized the pattern average, which did not vary (all patterns featured 12 black and 24 white cells). We used a fast Fourier transform algorithm (Cochran et al., 1967) for this purpose.

**Exclusion Criteria.** To identify participants who should be excluded from statistical analyses, we administered the following *seriousness check* (Aust et al., 2013): "It would be very helpful if you could tell us at this point whether you have taken part in this experiment seriously, so that we can use your answers for our scientific analysis, or whether you were just clicking through to take a look at the survey and did not rate the patterns seriously?". The response options were "I have taken part seriously" and "I have not taken part seriously, please throw my data away." All participants confirmed that they had taken part seriously, and no exclusions were therefore made.

**Additional Exploratory Variables and Demographics.** We assessed *political orientation* (liberal vs. conservative; Graham et al., 2009), *boredom proneness* using the short boredom proneness scale (Struk et al., 2017); *BIG 5 personality traits* using the ten-item personality inventory (TIPI; Gosling et al., 2003); *open-minded thinking* using the actively open-minded thinking scale (Haran et al., 2013); and *need for closure* using the brief 15-item need for closure scale (Roets & Van Hiel, 2011). As demographics, participants reported *age*, *gender*, and *nationality* in open-ended format.

## Results

### *Preliminary Analyses*

**Missing Data.** One participant did not provide complexity ratings for two patterns, randomness ratings for two patterns, and beauty ratings for one pattern. Data concerning these patterns for the participant were therefore missing and were not used in statistical analyses.

### *Main Hypothesis Testing*

**Subjective Indicators of Complexity and Randomness.** We first tested whether subjective randomness and complexity interacted in influencing perceived beauty (Hypothesis 1). We used multilevel modelling (Hayes, 2006) given that each participant provided multiple ratings, and complexity and randomness (level-1 predictors) were therefore nested under individual participants (level-2 variable).[3] To compute the models, we used the *nlme* package (Pinheiro et al., 2020) in R with Maximum likelihood (ML) estimation. We fit

[3] In all our analyses in the present study that employed multilevel models, we nested random slopes and intercepts for the predictor variables (level-1) under participants (level-2) but not under rating procedure (level-3) because the latter variable has only two levels and it would therefore not be optimal to use it as part of the nested structure (Finch et al., 2019; Hayes, 2006), but also because we wanted to specifically test whether rating scale would interact with the predictors (i.e., whether the link between the predictors and dependent variables differs depending on rating scale).

a random slopes model (i.e., with slopes and intercepts for each predictor and their interaction treated as random) rather than a random intercepts model (i.e., with only the intercepts treated as random) because comparing the fit of the two models using *anova* function in R showed that the former model had a better fit, $X^2(9) = 1923.513$, $p < .001$. All variables were *z*-standardized (Lorah, 2018).

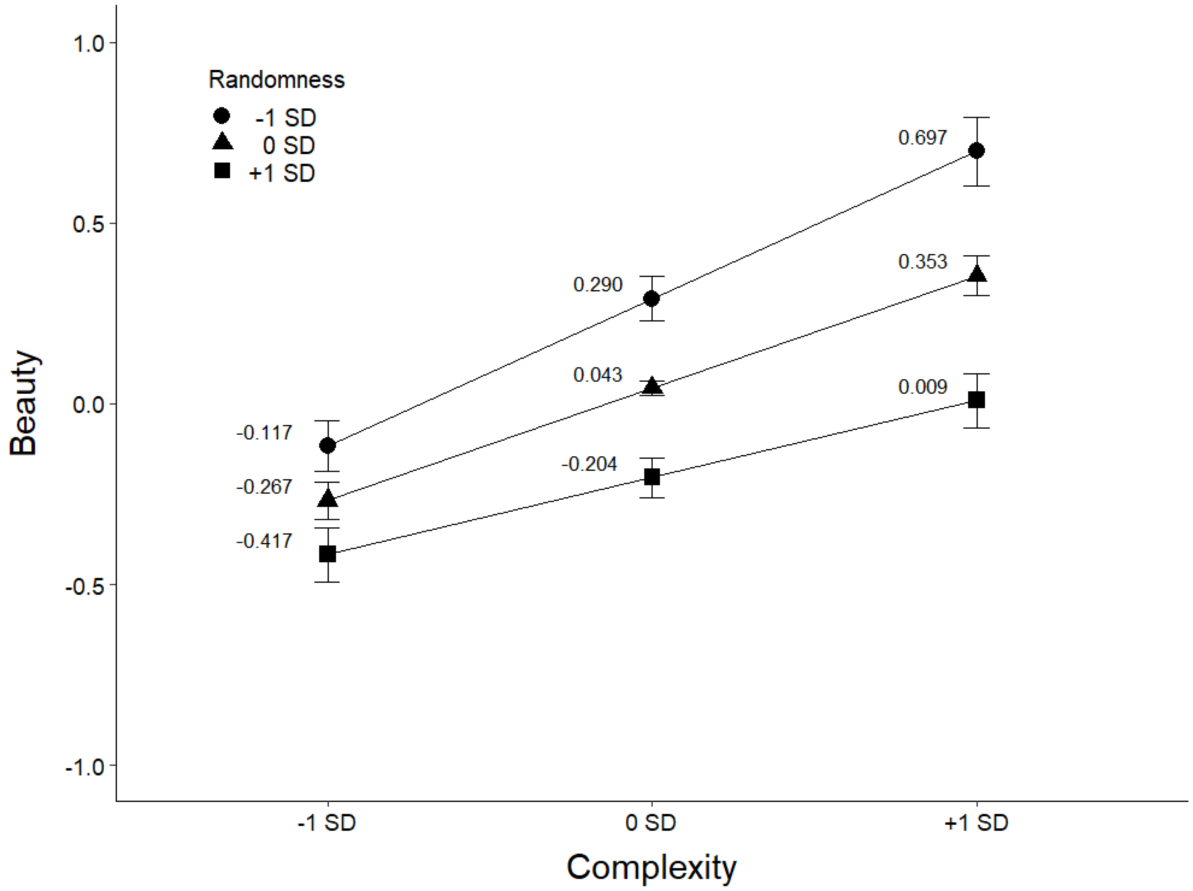


*Figure 6.* Graphical depiction of the interaction between subjective complexity and subjective randomness in predicting beauty (Study 1). Error bars correspond to the 95% Confidence Intervals.

The interaction between complexity and randomness in predicting beauty was significant, $b = -0.097$, 95% CI [-0.120, -0.074], $t(8486) = -8.368$, $p < .001$ (Figure 6). The main effects of complexity, *b* = 0.310, 95% CI [0.261, 0.359], *t*(8486) = 12.332, *p* < .001, and randomness, *b* = -0.247, 95% CI [-0.302, -0.192], *t*(8486) = -8.836, *p* < .001, were also significant. To further disentangle the pattern of the interaction, we performed the analysis of simple slopes (Bauer & Curran, 2005; Finch, Bolin, & Kelley, 2019). At high levels of randomness (+1 *SD*), complexity was positively related to beauty, *b* = 0.212, 95% CI [0.160, 0.264], *t*(8486) = 7.945, *p* < .001, whereas at the low levels (-1 *SD*) the relationship was also positive but roughly twice larger in magnitude, *b* = 0.408, 95% CI 0.351, 0.464], *t*(8486) = 14.143, *p* < .001 (Figure 6). Therefore, Hypothesis 1 was supported, given that, in line with our predictions, complexity and randomness jointly predicted beauty judgments, and most beautiful patterns tended to be those of low randomness and high complexity. These findings did not change depending on the rating procedure used (number vs. slider; Supplementary Materials, p.9).

**Objective Indicators of Complexity and Randomness.** We next tested whether objective randomness (Fourier randomness) and complexity (turns) would interact in influencing perceived beauty in line with predictions. We again fit a random slopes model (i.e., with slopes and intercepts for each predictor and their interaction treated as random) rather than a random intercepts model (i.e., with only the intercepts treated as random) because comparing the two models using *anova* function in R showed that the former model had a better fit, $X^2(9) = 522.697$, p < .001. As before, all variables in the model were *z*-standardized (Lorah, 2018).

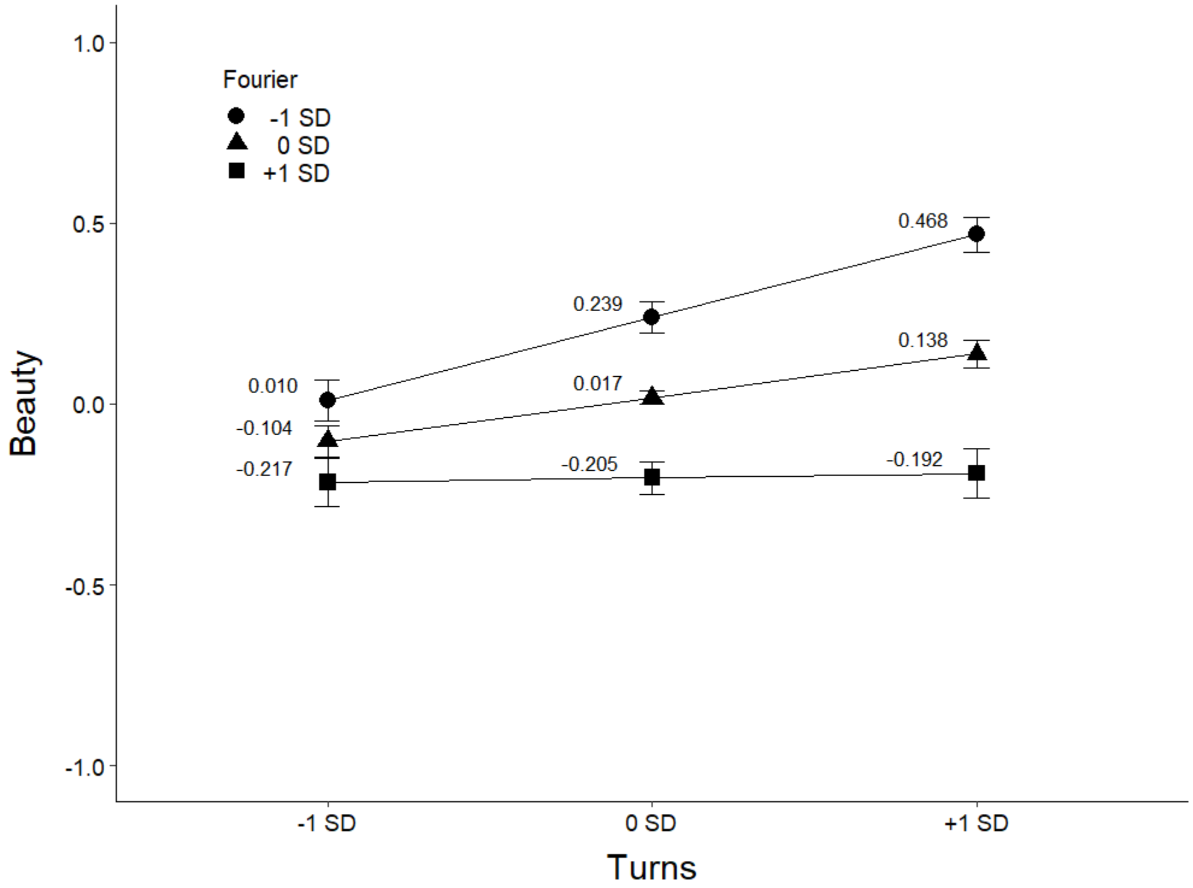


*Figure 7*. Graphical depiction of the interaction between objective complexity (Turns) and objective randomness (Fourier) in predicting beauty (Study 1). Error bars correspond to the 95% Confidence Intervals.

The interaction between turns and Fourier randomness in predicting beauty was significant, *b* = -0.108, 95% CI [-0.130, -0.086], *t*(8488) = -9.625, *p* < .001 (Figure 7). The main effects of turns, *b* = 0.121, 95% CI [0.085, 0.157], *t*(8488) = 6.630, *p* < .001, and Fourier randomness, *b* = -0.222, 95% CI [-0.261, -0.182], *t*(8488) = -11.040, *p* < .001, were also significant. To further disentangle the pattern of the interaction, we performed the analysis of simple slopes (Bauer & Curran, 2005; Finch et al., 2019). At high levels of Fourier randomness (+1 *SD*), turns was not related to beauty, *b* = 0.013, 95% CI [-0.038, 0.063], *t*(8488) = 0.491, *p* = .623, whereas at the low levels (-1 *SD*) the relationship was positive, $b = 0.229$, 95% CI [0.198, 0.260], $t(8488) = 14.532$, $p < .001$ (Figure 7). Therefore, the hypothesis was also supported for the objective measures, given that, in line with our predictions, objective complexity (turns) and objective randomness (Fourier) jointly predicted beauty judgments, and most beautiful patterns tended to be those of low objective randomness and high objective complexity. These findings remained robust regardless of the rating procedure used (number vs. slider; Supplementary Materials, p.9).

### *Additional Analyses*

**Comparing Current Pattern Complexity Ratings to Chipman (1977).** To test whether participants' complexity ratings from our study and Chipman (1977) were similar, we computed zero-order correlations between the two complexity variables averaged for each pattern. We were not able to compute multilevel models given that the data we obtained from Chipman (1977) contained averaged complexity ratings per each pattern. The correlation between subjective complexity in the present study and Chipman (1977) was high, $r(43) = .891$, thus indicating that the two variables were almost identical.

**Comparisons with Alternative Models and Exploratory Analyses.** Considering that previous theorizing and research on the link between complexity and beauty (e.g., Berlyne, 1970; Friedenberg & Liby, 2016; Güçlütürk et al., 2016; Nadal et al., 2010) demonstrated that complexity best predicts beauty through an inverted-U (i.e., quadratic) relationship, we compared quadratic models with the models we used in hypothesis testing, in which complexity predicted beauty through its interaction with randomness. Overall, the analyses showed that the models used in hypothesis testing were better predictors than quadratic models (Supplementary Materials, p.10). Moreover, in exploratory analyses, we tested whether political orientation, boredom proneness, each of the BIG 5 personality traits, open-minded thinking, and need for closure would moderate the interactions between *subjective* and *objective* complexity and randomness in predicting perceived beauty. Overall, these analyses showed that the interaction between complexity and randomness in predicting beauty was further moderated by open mindedness (Haran et al., 2013). That is, although this interaction was significant at all levels of open mindedness, it was stronger at higher relative to lower levels (for details, see Supplementary Materials, pp.10-11).

**Discussion**

Study 1 supported Hypothesis 1: complexity and randomness interacted in predicting beauty, and the most beautiful patterns were generally the ones with low randomness and high complexity. This finding was obtained when either subjective (i.e., perceived) complexity and randomness were used, or when their objective indicators (i.e., turns and Fourier randomness) were employed. Across all analyses, we found that the results did not differ depending on the pattern rating procedures (number versus slider). Importantly, we showed that the complexity ratings from our study were almost identical to the ratings from Chipman (1977), with the correlation effect size *r* being .891. This indicates that pattern perception is highly robust and is not dependent on a particular mode of presentation or pattern size, in line with what Chipman (1977) also suggested. Overall, although the present study provided a convincing support for Hypothesis 1, its main weakness is that it focused on a specific set of patterns (Chipman, 1977), and hence it remains possible that the hypothesis does not generalize across different possible black and white binary patterns. This weakness was addressed in the next study.

## Study 2

The previous study showed that complexity and randomness, quantified using objective indexes and their subjectively equivalents, interact in their relationship with aesthetic judgement. Specifically, patterns that combined high complexity with low randomness proved most aesthetically pleasing. Study 2 added two important improvements over the previous experiments: one theoretical, the other methodological.

As for the theoretical improvement, we examined more closely the psychological process that might link aesthetic judgement to complexity and randomness: quality attributions, which combine perceived skill and creativity. In Study 2, we therefore additionally tested if a complexity and randomness interaction emerged on quality attributions (Hypothesis 2), if quality attributions and beauty judgments shared a positive association (Hypothesis 3), and if the interaction effect of complexity and randomness on beauty judgements was mediated by attributed quality (Hypothesis 4).

The methodological improvement concerned our stimulus set. While the previous experiments offered insight into the interactive role of complexity and randomness in producing beauty, there is an important limitation to these studies: it is unclear if the stimuli used are a fair representation of these qualities in the stimulus-population that they must represent. Furthermore, extending the repertoire of stimuli benefits generalizability of the results beyond this very specific set of images.

Study 2, and the following studies also, addressed this issue by relying on stimuli that were more representative of the stimulus population they intended to represent at combinations of high and low randomness and complexity. First, we operationalized these high and low levels of complexity and randomness as their upper and lower tertiles in the Chipman (1977) set. We then generated new stimuli in a stepwise process inspired by mutation and selection processes in biological evolution: (1) we drew a random pattern; (2) we created a ‘mutation’ by randomly swapping the position of two cells; (3) we computed randomness and complexity for both original and mutation; (4) the pattern scoring closest to the target complexity and randomness (their tertile cutoffs) was retained and entered as original pattern in step (2). We repeated this process until we had 40 satisfactory patterns, 10 for each combination of high and low complexity and randomness. We then experimentally varied randomness and complexity in a within-factorial design.

## Method

### *Stimuli*

We generated 40 patterns that combined low and high randomness and low and high complexity following a 2 × 2 design; 10 patterns represented each of the four combinations (Figure 8). Patterns were 6 × 6 binary matrices with 12 black and 24 white elements. The 6 × 6 matrices were surrounded by a light-grey band of width identical to that of a single element. We operationalized low and high complexity as patterns containing fewer than 17 and more than 29 turns, respectively. We operationalized low and high randomness as patterns producing Fourier randomness below 17 or above 27.

Generation of each of these patterns for each of the four combinations followed a staged process: first, we randomly generated a pattern and calculated its number of turns and Fourier randomness. If corresponding complexity and randomness did not satisfy the set criteria (e.g., fewer than 17 turns and a Fourier amplitude count over 29)—which they invariantly did not—then the pattern entered a second and iterative stage. Here, a random pair of cells was selected, and their positions swapped. Complexity and randomness of the original pattern and its ‘mutation’ were then compared. If the mutation more closely satisfied the complexity and randomness criteria than its original without worsening on the other criterion, then it replaced the original in a next iteration of the same process. If the original and mutated pattern matched randomness and complexity, then one of these was retained at random. This iterative process of mutation and selection continued until (a) complexity and randomness criteria were met, or (b) a set maximum number of iterations was reached, at which point the entire process restarted with a new randomly assembled pattern. This process thus did not involve human interference with individual patterns except for setting the general criteria that patterns should adhere to.

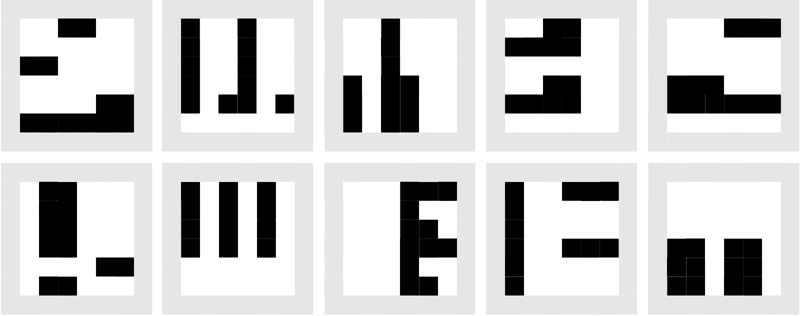
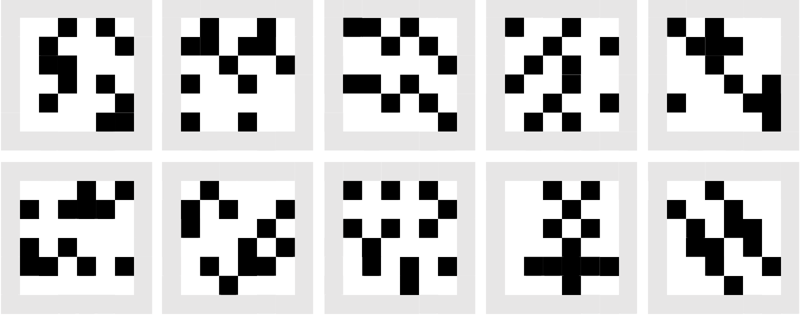
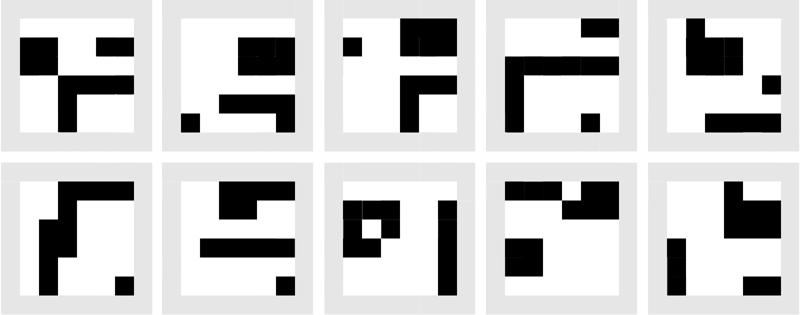
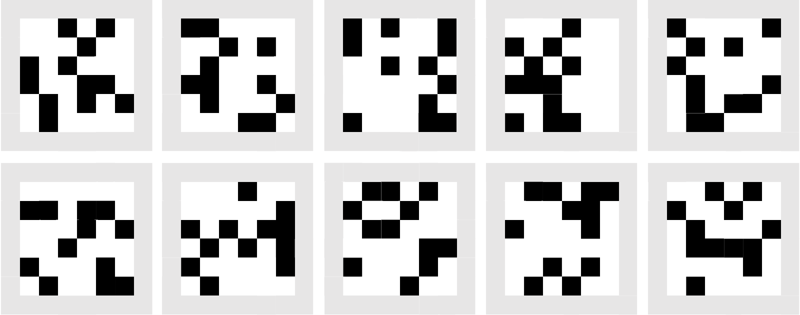
| Complexity | Randomness | Patterns |
|---|---|---|
| Low | Low |  |
| High | Low |  |
| Low | High |  |
| High | High |  |

*Figure 8*. Patterns used in Studies 2 through 4**.**

### *Determining sample size*

The novelty of stimuli and within-subject factorial design prevented us from having firm expectations of effect sizes. Therefore, we aimed for a sample large enough to detect a generic medium sized 2 × 2 within-subjects interaction effect (Cohen's $f = 0.10$) with a power of $(1 - \beta) = .80$, assuming moderate correlations between within-subject observations ($\rho$

= .50). The corresponding required sample size was *N* = 138 (Faul, Erdfelder, Lang, & Buchner, 2007), which we increased to 200 as a precaution against exclusions (see *Measures* section).

### *Participants and design*

Participants were 200 UK residents (Female = 139, Male = 61; $M_{\text{age}}$ = 33.145; $SD_{\text{age}}$ = 11.115), recruited at Prolific.co and paid £2.15 each. Participants underwent all conditions of the 2 (complexity: low, high) × 2 (randomness: low, high) within-subjects design. Application of exclusion criteria (see *Measures*) resulted in a final sample of 168 participants (119 women, 49 men; $M_{\text{age}}$ = 33.13, $SD_{\text{age}}$ = 10.97).

### *Procedure*

Participants completed the study online through Qualtrics. After giving consent, they reported demographics and received detailed task instructions. Specifically, they had to evaluate the 40 black and white patterns used as stimuli (Figure 8) in terms of their complexity, beauty, randomness, boredom, positivity, negativity, business, intensity, creativity, and skill (see Study 1). All 40 patterns were presented to participants together in a randomized order (i.e., they were not blocked according to complexity and randomness levels to ensure that participants could not easily infer our predictions). Furthermore, as in Study 1, participants were randomly allocated to one of the two quality rating procedures: assigning a relative number to each pattern (Chipman, 1977) or rating it on a slider from 0 (low) to 100 (high).

After evaluating the 40 patterns, participants completed exploratory individual difference measures. Among their items we included three attention checks where participants were asked to select a specific value on a scale. Finally, participants received a seriousness check where they could confirm if their data should be included (Aust et al., 2013), and were then debriefed.

### *Measures*

**Dependent variable.** Participants' average ranks for the patterns' beauty served as dependent variable. As in Study 1, participant's beauty ratings of all 40 patterns were first ranked (e.g., Chipman, 1977). We then calculated average ranks for each of the four sets of 10 patterns, representing the four combinations of complexity and randomness. Higher average ranks indicate greater beauty.

**Mediators.** Our candidate mediator, quality, was measured through assessing perceived creativity and perceived skill (Kozbelt, 2004). These two elements were rated as part of the pattern evaluations. The preceding instructions informed participants that ratings of "creativity" and "skill" referred to "how creative a pattern is" and "how much skill it takes to create the pattern", respectively. As for beauty ratings, we calculated average creativity and skill ranks for each of the four sets of patterns. Skill and creativity were highly correlated with each other in each of the four conditions (*r*s ≥ .718, *p*s < .001) and were averaged into an index of quality accordingly.

**Manipulation checks: Subjective complexity and randomness.** We manipulated objective complexity and randomness by directly altering the composition of stimuli patterns. We used participants' ratings of (subjective) complexity and randomness to verify that this manipulation of complexity and randomness corresponded to their subjective equivalents. As we did for beauty, ratings were first ranked, and we then computed averages for each condition. Higher ranks indicate higher subjective complexity and randomness.

**Exclusion criteria.** We attempted to identify participants who did not pay attention to the study content with two methods. First, we included three *instructed-response check* items that asked participants to select a specific value on an interval scale (e.g., Please select "Strongly agree"; Kung, Kwok, & Brown, 2018; Meade & Craig, 2012; Thomas & Clifford, 2017). These checks were placed among the various items of the exploratory personality measures (see *Additional Exploratory Variables*); failing to answer them correctly led to exclusion. Second, we administered a seriousness check where participants were invited to self-disclose if their data should be excluded from analyses (Aust et al., 2013) as in Study 1. All participants who did not correctly answer all check items (i.e., the three instructed-response checks and the seriousness check) were excluded from statistical analyses (n = 32).

**Additional Exploratory Variables and Demographics.** As in Study 1, we included several measures to explore moderation by their corresponding constructs. These included *political orientation* (liberal vs. conservative; Graham et al., 2009), *boredom proneness* (Struk et al., 2017); the *BIG 5 personality traits* (Gosling et al., 2003); *open-minded thinking* (Haran et al., 2013); and *need for closure* (Roets & Van Hiel, 2011). As demographics, participants reported their *age, gender,* and *nationality* in open-ended format.

## Results

### *Preliminary Analyses*

**Missing Data.** One participant had missing values on six or more ratings for each evaluated feature, in each condition. We excluded data for this participant in statistical analyses as key variables could not be reliably computed from these data.

**Manipulation Checks.** We first examined if the objective differences in high versus low complexity, and high versus low randomness received corresponding subjective ratings on these constructs. We tested this by entering subjective complexity and subjective randomness as dependent variables in two within-subjects ANOVAs, with manipulated (objective) complexity and randomness as independent variables.

Regarding subjective complexity, we found a significant and very large main effect of the complexity manipulation, $F(1, 166) = 778.255$, $p < .001$, $\eta_p^2 = .824$, as well as a small main effect of the randomness manipulation, $F(1, 166) = 6.039$, $p = .015$, $\eta_p^2 = .035$. We also found a significant complexity × randomness interaction, $F(1, 166) = 62.121$, $p < .001$, $\eta_p^2$

= .272 (Figure 9). This interaction suggested that the magnitude of the impact of the complexity manipulation on subjective complexity varied somewhat across low and high randomness conditions. Contrast analysis confirmed that the difference between low and high complexity conditions was nonetheless significant, and substantial, in both the low randomness, $M_{\text{diff}}$ = 14.738, $p$ < .001, 95% *CI* [13.673, 15.804], $\eta_p^2$ = .818, and in the high randomness, $M_{\text{diff}}$ = 11.152, $p$ < .001, 95% *CI* [10.179, 12.124], $\eta_p^2$ = .754, condition.

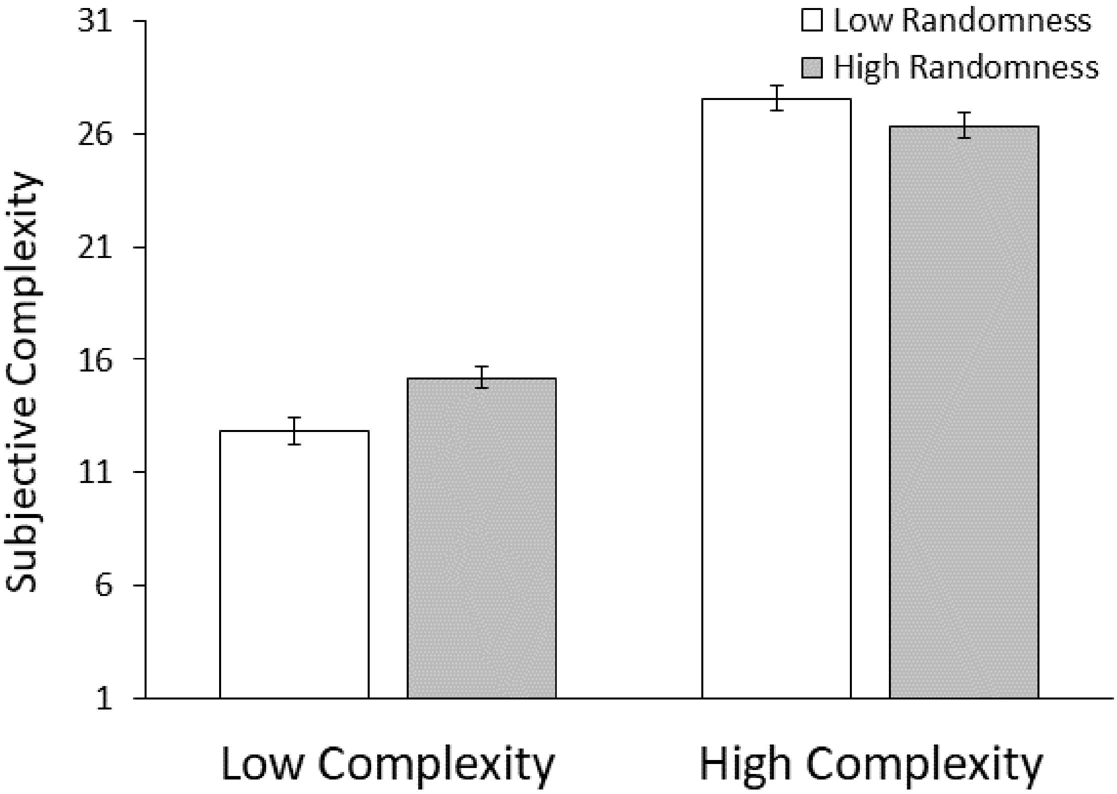


*Figure 9.* Subjective complexity as a function of objective complexity and objective randomness (Experiment 2). Error bars correspond to the 95% Confidence Intervals.

A similar analysis for subjective randomness confirmed a significant and substantial main effect of the randomness manipulation, $F$(1, 166) = 135.353, $p$ < .001, $\eta_p^2$ = .449. Also, the main effect of the complexity manipulation was significant, $F$(1, 166) = 125.795, $p$ < .001, $\eta_p^2$ = .431, and the complexity × randomness interaction was not, $F$(1, 166) = 5.792, $p$

= .496, $\eta_p^2$ = .003 (Figure 10). These results show that the manipulations of complexity and randomness were successful; objective differences in them transferred to corresponding subjective perceptions.[4]

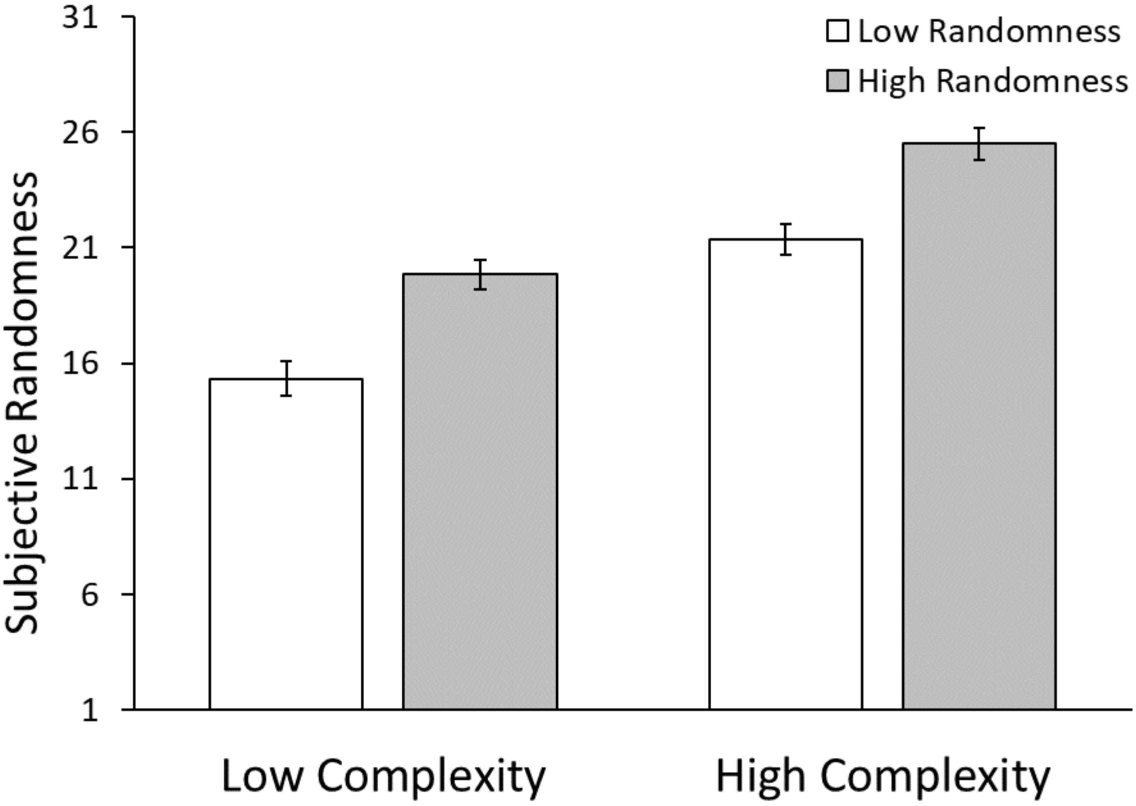


*Figure 10.* Subjective randomness as a function of objective complexity and objective randomness (Experiment 2). Error bars correspond to the 95% Confidence Intervals.

[4] Results indicated some cross-over in the subjective perception in the sense that subjective complexity and randomness were not as neatly separated as their objective equivalents. A critical perspective might argue that, perhaps, complexity or randomness are either subjectively indistinct, or that one might subsume the other. To verify that (1) the complexity manipulation altered subjective complexity above and beyond changes in subjective randomness, and (2) that the randomness manipulation altered subjective randomness above and beyond subjective complexity, we reran our analyses with either subjective randomness or subjective complexity as covariate in a set of maximum likelihood random-intercept multilevel regressions. Results confirmed that (1) objective complexity still increased subjective complexity after controlling for subjective randomness, $B$ = 5.976, $SE$ = 0.166, $t$(497) = 35.990, $p$ < .001, 95% CI [5.651, 6.301], and (2) objective randomness still increased subjective randomness after controlling for subjective complexity, $B$ = 2.088, $SE$ = 0.171, $t$(497) = 12.222, $p$ < .001, 95% CI [1.753, 2.422].

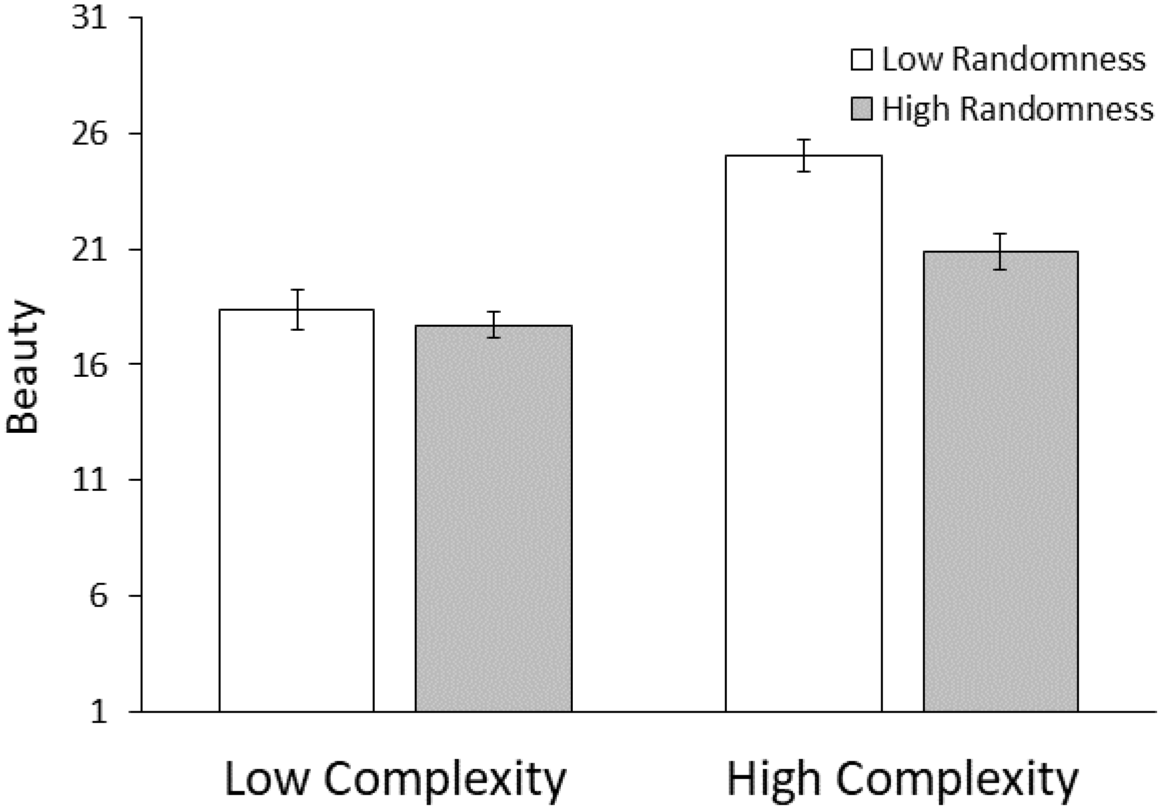


*Figure 11*. Beauty judgements as a function of objective complexity and objective randomness (Study 2). Error bars correspond to the 95% Confidence Intervals.

### *Main Hypothesis Testing*

**Beauty combines high complexity with low randomness.** We entered beauty as dependent variable in a 2 (complexity: high, low) × 2 (randomness: high, low) within-subjects ANOVA. This analysis produced main effects of both manipulated complexity, $F(1, 166) = 62.601$, $p < .001$, $\eta_p^2 = .274$, and manipulated randomness, $F(1, 166) = 49.361$, $p < .001$, $\eta_p^2 = .229$. Critically, we also found a significant interaction with a considerable effect size (for an interaction), $F(1, 166) = 44.291$, $p < .001$, $\eta_p^2 = .211$ (Figure 11). Contrast analyses indicated that the increase in beauty that low vs. high manipulated randomness caused was greater for patterns high in complexity, $M_{\text{diff}} = 4.154$, $p < .001$, 95% *CI* [3.275, 5.033], $\eta_p^2$ = .344, than low in complexity, $M_{\text{diff}}$ = 0.699, *p* = .097, 95% *CI* [-0.127, 1.525], $\eta_p^2$ = .017. These results confirm that, as hypothesized, the combination of high complexity with low randomness renders patterns particularly attractive. The findings remained robust regardless of the rating procedure used (number vs. slider; Supplementary Materials, p.12).

**Mediation by quality.** So far, results confirmed that patterns that feature both high complexity and low randomness are perceived as comparatively beautiful. We next tested if participants attribute quality (i.e., creativity and skill) to such patterns. After all, attributions of creative skill act as a precursor to aesthetic judgement (Van Tilburg & Igou, 2014). We examined this with a statistical mediation approach (Hayes, 2009), where we tested (I) if patterns that combine high complexity with low randomness are seen as particularly high in quality, (II) if perceived quality predicts beauty judgements after controlling for complexity, randomness, and their interaction, and (III) if an indirect effect can be traced from the complexity × randomness interaction on beauty judgements through perceived quality. Step (II) of this analysis requires coefficient estimation for categorical (complexity, randomness) and continuous (quality) statistical predictors; we accommodated this by relying on maximum likelihood multilevel regression models throughout, where the four conditions and their corresponding evaluations were nested within participants.

Perceived quality was regressed on (effect coded) complexity condition (-1 = low, 1 = high), the randomness condition (-1 = low, 1 = high), and the complexity × randomness interaction. These predictors were treated as fixed variables and participants were assigned a random intercept—resulting in a random-intercept multilevel regression model. This analysis returned significant main effect of complexity, *B* = 3.599, *SE* = .162, *t*(498) = 22.174, *p* < .001, 95% *CI* = [3.281, 3.917], randomness, *B* = -.383, *SE* = .162, *t*(498) = 2.358, *p* = .019, 95% *CI* = -0.701, -0.065], and, importantly, their interaction, *B* = -.845, *SE* = .162, *t*(498) = 5.235, *p* < .001, 95% *CI* = [-1.168, -.532]. These results show that especially patterns combining high complexity with low randomness were attributed quality (indeed, all *p*s < .001), supporting (I).

We next ran a similar random-intercept analysis in which perceived beauty was regressed on complexity, randomness, their interaction, and quality. This analysis returned a significant partial effect of randomness, *B* = -.920, *SE* = .143, *t*(497) = 6.421, *p* < .001, 95% *CI* = [-1.200, -.639], no significant partial effect of complexity, *B* = -0.300, *SE* = .188, *t*(497) = 1.595, *p* = .111, 95% *CI* = [-.669, .068], and no significant complexity × randomness interaction, *B* = -0.213, *SE* = .146, *t*(497) = 1.450, *p* = .145, 95% *CI* = [-.497, .072]. Critically, the partial association between quality and beauty was significantly positive, *B* = .767, *SE* = .034, *t*(497) = 22.469, *p* < .001, 95% *CI* = [.700, .833]. These results suggest that the initially significant interaction effect of complexity and randomness on beauty was rendered mute after controlling for quality; quality, in turn, replaced it as significant predictor of beauty perceptions, supporting (II).

We next tested if the indirect effect of the complexity × randomness interaction on beauty through quality was significant (note that this effect is equivalent to the change in the complexity × randomness interaction effect by including quality). We tested this using the Monte-Carlo estimation tool by Selig and Preacher (2008, 20,000 repetitions). This analysis revealed that the indirect effect (*B* = -0.648) was indeed significant, 95% *CI* = [-.904, -.401]. The effect of the complexity × randomness interaction on perceived beauty was significantly mediated by quality, confirming (III).

### ***Exploratory Analyses***

We probed if additional exploratory variables (see the *Measures* section) moderated the complexity × randomness interaction on beauty judgements. We ran a separate analysis for each putative moderator (9 in total). As the significance level, we adopted .006 (i.e., 0.05

divided by the number of analyses conducted) and used random-intercept multilevel models with ML estimation. These analyses produced no significant triple interactions (all *p*s $\geq .062$).

## Discussion

Study 2 supported the aesthetic quality model on a set of black and white patterns drawn from a representative population of these stimuli comprising different combinations of low and high objective complexity (turns) and randomness (Fourier). In line with Hypothesis 1, complexity and randomness interacted in influencing beauty judgments: the most beautiful patterns were the ones with low randomness and high complexity. As predicted by Hypothesis 2, this interaction also influenced the proposed mechanism: quality attributions (i.e., creativity combined with skill). Highest quality was attributed to non-random but highly complex patterns. Moreover, quality was positively associated with beauty judgments (Hypothesis 3), and hence the interaction effect of complexity and randomness on beauty judgements was mediated by this variable (Hypothesis 4). Overall, although Study 2 comprehensively supported the aesthetic quality model on a representative set of patterns and thus produced generalizable findings (Westfall et al., 2015), its main limitation is that we did not causally manipulate the proposed mechanism (e.g., Spencer et al., 2005). This limitation was addressed in the next studies.

# Study 3

In the previous study, we showed that quality statistically mediated the interactive impact of pattern randomness and complexity on beauty judgements. A weakness of demonstrating the mechanism using this statistical approach is that the link between quality and the dependent variable is correlational, and hence it remains possible that some other “true” mediator may in fact drive the effect of the patterns on perceived beauty (Pirlott & MacKinnon, 2016; Spencer et al., 2005).

To address this issue, in the present study we experimentally manipulated the mediator by emphasizing versus undermining pattern quality. We focused on the high complexity patterns (low and high in randomness) adopted from Study 2. Specifically, in one condition we told participants that the complex non-random patterns (i.e., the more beautiful patterns) were created by a graphic designer with the aim to be creative and imaginative (i.e., of high quality), whereas the complex random patterns (i.e., the less beautiful ones) were created by a computer with the aim to be uncreative and unimaginative (i.e., of low quality). We refer to this condition as ‘compatible’ because the quality we attributed to patterns matched participants’ actual perception assessed in Study 2. In contrast, in the ‘incompatible’ condition, participants were told that the complex random patterns were created by a graphic designer to be creative and imaginative, whereas the complex non-random patterns were created by a computer to be uncreative and unimaginative. We included also a third condition, called ‘neutral’, in which we did not provide any description regarding how the patterns were created.

Consistent with the moderation-of-process approach to test causal mediation (Spencer et al., 2005), we predicted that if quality attribution indeed accounts for the effects of complex non-random versus random patterns on beauty, then the difference in perceived beauty between the two types of patterns should be smaller in the incompatible condition compared to either the compatible or the neutral condition. Indeed, we expected this because pairing the less (vs. more) beautiful patterns with high (vs. low) quality should elevate (vs. reduce) the patterns’ beauty ratings, thus making the difference between them smaller. We did not have a specific prediction regarding the compatible compared to neutral condition.

Finally, it is important to clarify why in Study 3, in addition to the complex non-random patterns (i.e., the most beautiful ones) that are of key interest for our aesthetic quality model, we tested only the complex random patterns, but not the less complex ones. In the context of experimentally assessing a mechanism, there is a limit to what can be manipulated (Spencer et al., 2005). More specifically, in Study 3 we relied on priming to demonstrate the mechanism, given that we induced the mental constructs of low versus high creativity and expected this “knowledge activation” would change beauty judgments (Bargh, 2006; Gawronski & Bodenhausen, 2005; Rietzschel, Nijstad, & Stroebe, 2007). It is well known that priming as an experimental technique has several limitations (e.g., Cesario, 2014; Ramscar, 2016), and it is more likely to work when its intended effect on perceptions or behavior is not fully at odds with participants’ underlying beliefs and motives (e.g., Shariff, Willard, Andersen, & Norenzayan, 2016; Van Koningsbruggen, Stroebe, & Aarts, 2011). In our Study 2, representative non-complex patterns (either random or non-random) were judged as less beautiful than either of the two complex pattern types (all $p$s $\leq .002$). Given that participants therefore generally perceived the non-complex patterns to be low in beauty, we were skeptical that it would be possible to prime people to perceive these patterns as more beautiful because this would be too inconsistent with their actual beliefs. Indeed, we assumed that, due to the limitations of priming as a technique, it would be more optimal to test the complex random patterns: even if these stimuli are judged as less beautiful than the complex non-random ones, they are more beautiful than the non-complex patterns and experimentally increasing their beauty via primed quality would be more plausible due to a smaller incompatibility with people’s actual beliefs.

## Method

### *Stimuli*

The stimuli in the present study were the 20 high complexity patterns from Study 2, consisting of 10 patterns low in randomness and 10 high in randomness.

### *Determining Sample Size*

The present study was different from the previous ones and, therefore, we did not have a precise estimate about the expected effect size. Using G*Power (Faul et al., 2007), we computed the number of participants that need to be tested to detect a medium effect size (Cohen's $f$ = 0.25). *ANOVA: Fixed effects, omnibus, one-way* was selected, and a prior power analysis was implemented. As *power* we used .80, as significance level .05, and as the number of groups we inputted 3 (corresponding to the three conditions tested in the present study). The analysis indicated that 159 participants should be recruited. To be on the safe side and ensure that this sample size is met after applying the exclusion criteria (see the *Measures* section below), we tested 243 participants, which resulted in the final sample size of 182 participants included in statistical analyses (see *Participants and design*).

### *Participants and Design*

Two hundred and forty-three participants of UK nationality completed the online study (Female = 156, Male = 87; $M_{\text{age}}$ = 40.284; $SD_{\text{age}}$ = 12.494) via Prolific.co. Payment was £2.15. We used a 3-level between-subjects design with *compatibility* (incompatible vs. compatible vs. neutral) as the independent variable. After the exclusion criteria were applied (see the *Measures* section below), 182 participants were eventually included in statistical analyses (Female = 116, Male = 66; $M_{\text{age}}$ = 40.692; $SD_{\text{age}}$ = 12.640), thus leaving 64 participants in the incompatible condition, 52 in the compatible condition, and 66 in the neutral condition.

### *Procedure*

Participants gave consent and reported demographics (see the *Measures* section below). Then, they were randomly allocated to one of the three conditions and received corresponding general instructions. Participants in all conditions were told to score two different sets of black and white patterns (each consisting of 10 pattens) on various qualities (i.e., complexity, beauty, randomness, creativity, and skill). The scoring procedure and each quality were described as in the previous studies. Immediately before scoring the complex random patterns, participants in the incompatible condition were informed that these patterns were created by a graphic designer with the goal to be creative, whereas immediately before scoring the complex non-random patterns they were informed that these patterns were produced by a computer with the aim of them being uncreative and unimaginative. In the compatible condition, the instructions were reversed. Before scoring each of the two sets of patterns, participants in the incompatible and compatible conditions were given an understanding check item (see the *Measures* section below) to ensure they accurately recalled whether the patterns they were about to score were creative or uncreative. Participants in the neutral condition did not receive these additional instructions.

Then, participants evaluated the patterns on complexity, beauty, randomness, creativity, and skill. We used only continuous scoring with sliders because this procedure was less effortful and took less time compared to the number scoring, and our previous studies did not found differences between these methods. After participants evaluated the patterns, they received a general understanding check item (see the *Measures* section below), after which they filled in various exploratory individual-differences measures in which instructed-response items (see the *Measures* section below) were embedded to further identify participants who were not paying attention. Finally, they received a seriousness check (Aust et al., 2013) and were debriefed.

### *Measures*

**Dependent Variable.** The dependent variable was the *difference* in beauty between the (high complexity) low randomness and high randomness patterns. To compute this, we first transformed a participants' beauty ratings of all 20 patterns into ranks using the procedure from the previous studies, assigning an average rank to sets of duplicates. Higher ranks indicated higher beauty. Then, for the two subsets of 10 patterns we created an average score.

We then subtracted the score of the high randomness patterns from that of the low randomness patterns. Positive values thus indicated that participants perceived the low randomness patterns (vs. high randomness patterns) as more beautiful, whereas negative values indicated the opposite. We computed the dependent variable using this procedure because it allowed us to directly test our main prediction (that the difference in perceived beauty between non-random versus random patterns would be smaller in the incompatible compared to the compatible or neutral conditions) via a simple between-subjects ANOVA. We did not use the average beauty rankings themselves as the dependent variable because a 3 × 2 mixed ANOVA probing the influence of the interaction between condition and pattern randomness (high vs. low) on these rankings would allow us to understand only whether the differences between the two pattern types changed across conditions, but not to directly examine our prediction (i.e., whether the difference in the incompatible condition was smaller than in each of the other two conditions). This more elaborate analysis is, however, available in Supplementary Materials (pp.14-16).

**Manipulation Checks.** As manipulation checks, we averaged the differences in creativity and skill between high randomness and low randomness patterns, thus indicating the difference in quality between the two pattern types. This index was therefore computed using a similar procedure as the dependent variable. Positive values indicated that participants attributed higher quality to the low randomness patterns (vs. high randomness patterns), whereas negative values indicated the opposite.

**Exclusion Criteria.** We used several check items to identify participants who should be excluded from statistical analyses. Two *understanding check* items were administered to participants in the incompatible and compatible conditions and required them to confirm whether the patterns they were about to rate were creative or uncreative based on the instructions they received. Three response options were offered: “Uncreative”, “Creative”, and “I do not remember”. Moreover, all participants received a *general understanding check*, for which they had to confirm what the study was about among the following seven options: “Rating colorful patterns on dimensions such as complexity, creativity, beauty, etc.”; “Rating a combination of black and white and colorful patterns on dimensions such as complexity, creativity, beauty, etc.”; “Rating black and white patterns on dimensions such as complexity, creativity, beauty, etc.” (this was the correct answer); “Counting the number of black squares in black and white patterns”; “Interpreting figures in black and white patterns.”; “Counting the number of white squares in black and white patterns”; and “Indicating your preference for colorful patterns.” All participants responded to three *instructed-response check* items (e.g., Please select "Strongly agree" in response to this question; Kung et al., 2018; Meade & Craig, 2012; Thomas & Clifford, 2017). Finally, participants received the *seriousness check* (Aust et al., 2013) at the end of the study, as we did in the previous ones. All participants who did not correctly answer all check items (i.e., the two understanding check items, the general understanding check, the three instructed-response items, and the seriousness check) were excluded from statistical analyses (n = 61).

**Additional Exploratory Variables and Demographics.** As in the previous studies, we assessed *political orientation* (liberal vs. conservative; Graham et al., 2009). Moreover, we measured *boredom proneness* using a short boredom proneness scale (Struk et al., 2017); *BIG 5 personality traits* using the ten-item personality inventory (TIPI; Gosling et al.,2003); *open-minded thinking* using the actively open-minded thinking scale (Haran et al., 2013); and *need for closure* using the brief 15-item need for closure scale (Roets & Van Hiel, 2011). As demographics, participants reported their *age*, *gender,* and *nationality* as in Studies 1-2.

## Results

### *Preliminary Analyses*

**Missing Data.** Concerning beauty, one participant did not rate three (out of ten) random patterns. Concerning, creativity, this person did not rate six (out of ten) random patterns and one (out of ten) non-random patterns. Concerning skill, this participant did not provide scores for any of the random and non-random patterns. Moreover, concerning complexity, this person did not rate five (out of ten) random patterns. Finally, regarding randomness, this person did not rate one (out of ten) random patterns. Data for this single participant were therefore not included in statistical analyses because the main variables tested in this study (i.e., the dependent variable and the manipulation checks) could not be reliably computed.

**Pattern Randomness, Beauty, and Quality.** In the previous study, we showed that low-randomness patterns were perceived as more beautiful and judged as being of higher quality than high-randomness patterns. We verified if the same effects occurred in the present study. A repeated measures ANOVA with pattern randomness (high vs. low) as within-subjects variable, and beauty rank as dependent variable confirmed that patterns low in randomness ($M = 11.693$; $SD = 2.080$) were perceived as more beautiful than highly random patterns ($M = 9.307$; $SD = 2.080$), $F(1, 180) = 59.503$, $p < .001$, $\eta_p^2 = .248$. Another repeated measures ANOVA with quality rank as dependent variable showed that non-random patterns ($M = 11.506$; $SD = 2.314$) were judged to be of higher quality than random patterns ($M = 9.494$; $SD = 2.314$), $F(1, 180) = 34.255$, $p < .001$, $\eta_p^2 = .160$. Finally, quality was strongly correlated with perceived beauty for either random or non-random patterns, $r(179) = .687$, $p < .001$.[5] These analyses therefore supported the main assumptions of the aesthetic quality model.

[5] Correlation effect sizes for both complex and non-complex patterns were the same because of how the variables in question were computed (i.e., beauty and quality ranks were calculated for each participant across the random and non-random patterns this participant evaluated).

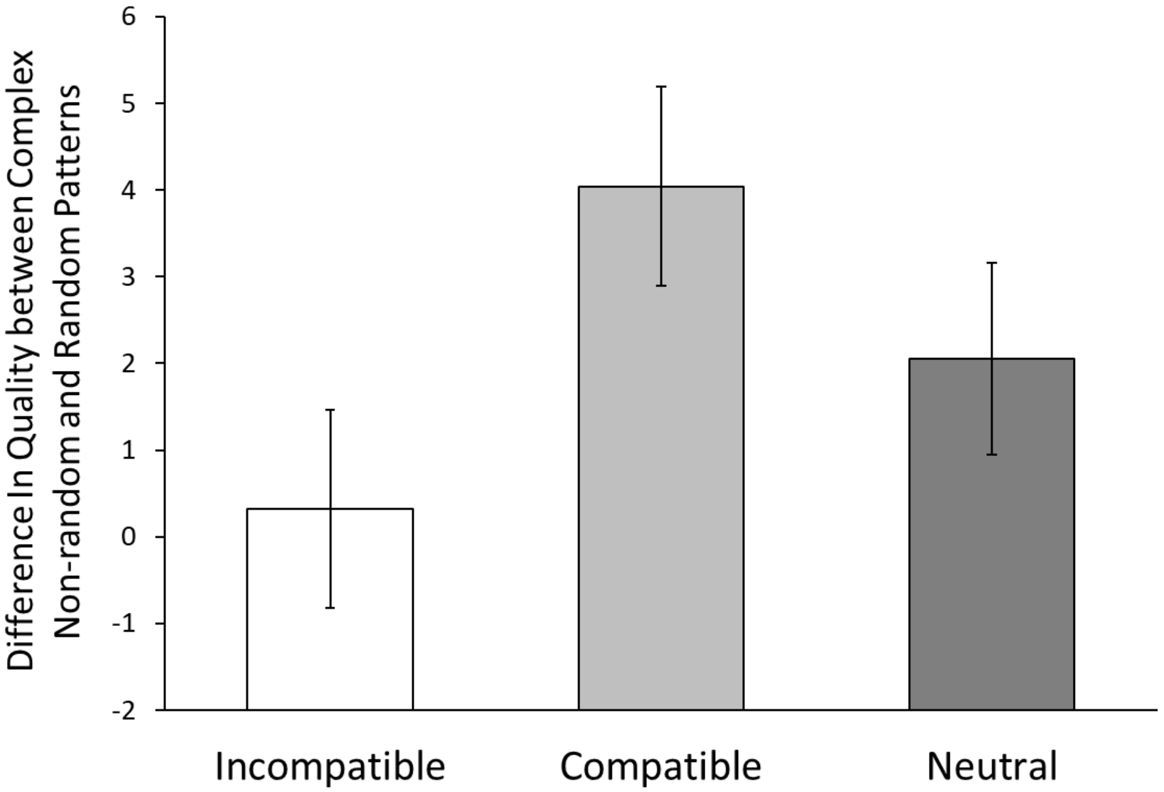


*Figure 12.* Manipulation check: the difference in attributed quality between complex non-random and random patterns as a function of the incompatible, compatible, and neutral conditions (Study 3). Positive values indicate that participants attributed higher quality to the non-random (relative to random) patterns. Error bars correspond to the 95% Confidence Intervals.

**Manipulation Check.** To test whether *compatibility* influenced differences in quality attributions for random vs. non-random patterns, we performed a one-way ANOVA. The result was highly significant (Figure 12): compatibility influenced the quality manipulation check, $F(2, 178) = 10.213$, $p < .001$, $\eta_p^2 = .103$. Planned contrasts further showed that the incompatible condition had a lower difference in quality between complex non-random and random patterns compared to both the compatible, $M_{\text{diff}} = 3.718$, $p < .001$, 95% CI [2.094, 5.341], and neutral condition, $M_{\text{diff}} = 1.733$, $p = .027$, 95% CI [0.201, 3.264], as predicted

(Figure 12). The difference between the compatible and neutral conditions for which we did not have a clear prediction was also significant, $M_{\text{diff}}$ = 1.985, $p$ = .016, 95% CI [0.367, 3.603] (Figure 12). Considering that in Study 3 we experimentally manipulated quality by focusing on creativity, a critic may argue that the study primarily provides evidence regarding creativity (rather than quality as a whole) as a mechanism. To address this criticism, in Supplementary Materials (pp.13-14) we report analyses for creativity and skill manipulation checks individually to show they were impacted by *compatibility* almost identically and were highly correlated, $r(179) = .865$, thus indicating that the manipulations we used tackled quality as a whole.

### ***Main Prediction: Compatibility and Differences in Beauty between Patterns***

A one-way ANOVA testing whether *compatibility* influenced the difference in beauty judgments between low randomness and high randomness patterns was significant, $F(2, 178) = 3.895$, $p$ = .022, $\eta_p^2$ = .042 (Figure 13). Planned contrasts further showed that the incompatible condition had a lower beauty difference score compared to both the compatible, $M_{\text{diff}}$ = 2.023, $p$ = .009, 95% CI [0.515, 3.532], and neutral condition, $M_{\text{diff}}$ = 1.470, $p$ = .043, 95% CI [0.046, 2.893], as predicted (Figure 13). The difference between the compatible and neutral conditions for which we had no prediction was not significant, $M_{\text{diff}}$ = 0.554, $p$ = .468, 95% CI [-0.950, 2.057]. Additional analyses probing how specific combinations of compatibility and pattern randomness impacted beauty judgments are available in Supplementary Materials (pp.14-16).

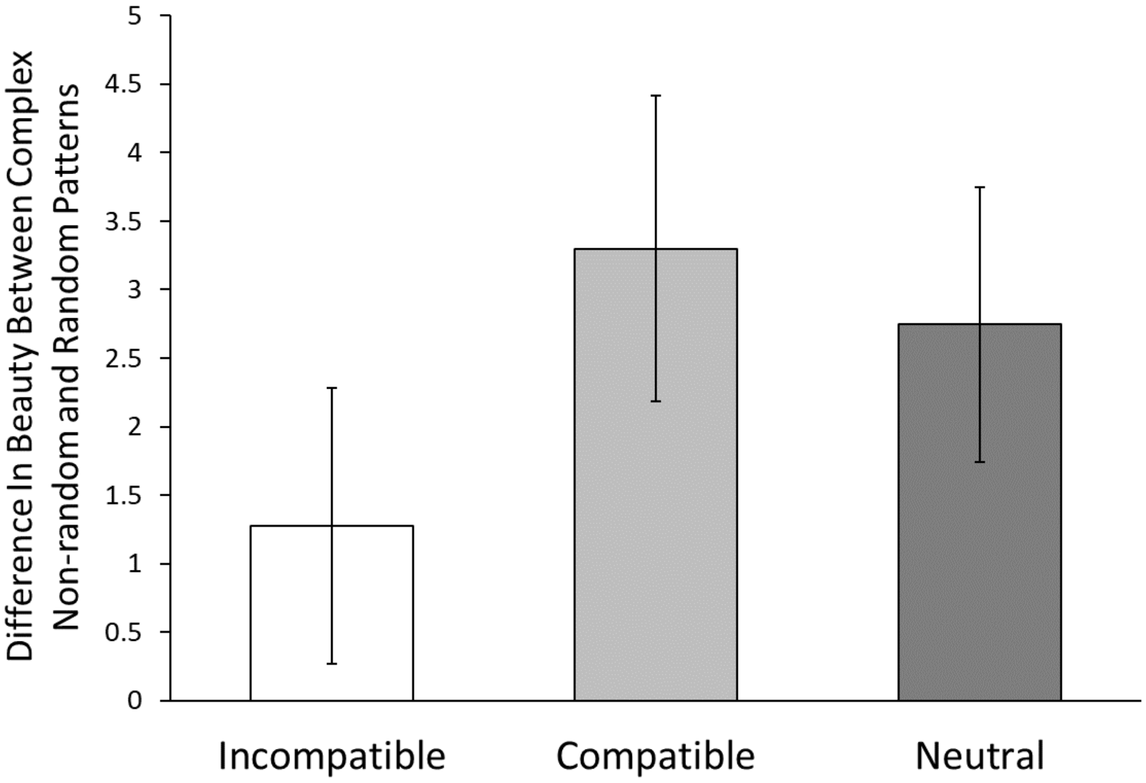


*Figure 13.* The difference in beauty between complex non-random and random patterns as a function of the incompatible, compatible, and neutral conditions (Study 3). Positive values indicate that participants attributed higher beauty to the non-random (relative to random) patterns. Error bars correspond to the 95% Confidence Intervals.

### ***Exploratory and Additional Analyses***

In exploratory analyses, we probed whether the individual differences measures tested (see the Measures section) would moderate the influence of *compatibility* on the dependent variable. The interaction between each variable and compatibility was computed in a separate multiple regression analysis. As the significance level, we used .006 (i.e., 0.05 divided by the number of moderators tested—nine). No interaction effects were significant, all *p*s ≥ .024. In Supplementary Materials (pp.16-17), we also report additional analyses testing the differences between the patterns we used as stimuli concerning subjective complexity and randomness.

## Discussion

Study 3 supported our main prediction: the incompatible (vs. neutral or compatible) condition decreased the difference in perceived beauty between complex non-random and random patterns. The manipulation checks further showed that our experimental intervention successfully manipulated quality that participants attributed to patterns: the difference between the non-random and random patterns in terms of quality was smallest in the incompatible (vs. neutral or compatible) condition, in congruence with the results for beauty.

A critical reading may ask if the results of the present study could be explained by demand characteristics. For example, we told participants that some patterns were creative versus uncreative, which might have signaled that we wanted them to rate these patterns as beautiful versus ugly. However, if this were the case, participants would have rated the creative patterns as more beautiful than the uncreative ones in both the incompatible and compatible conditions. That is, in the incompatible condition, random patterns associated with creativity would have been rated as more beautiful than the non-random patterns lacking creativity, whereas the effect would have reversed in the compatible condition. In contrast, as can be seen in Figure 13, participants in either of the two conditions judged non-random (vs. random) patterns to be more beautiful (i.e., the difference in creativity between non-random and random patterns was always positive), and it was only the relative difference between the patterns that changed across the conditions. It is therefore unlikely that demand characteristics can explain the present findings.[6]

[6] In the present research, we did not manipulate creativity as a between-subjects variable to understand how it independently impacts beauty judgments, and whether this impact is similar across non-random and random patterns, given that this was not one of the key questions our study aimed to assess. However, a visual inspection of supplementary Figures S2 and S4 (Supplementary Materials, pp.16 & 21) indicates that manipulating high vs. low creativity similarly increased beauty judgments for either the random or non-random patterns (the same finding was obtained in Studies 3 and 4). That is, non-random patterns associated with high creativity (compatible condition) vs. low creativity (incompatible condition) were judged as more beautiful, and this difference was comparable to the one between random patterns associated with high creativity (incompatible condition) vs. low creativity (compatible condition). In other words, patterns allegedly made by a human designer (i.e., the creative ones) were on average judged as being more beautiful than patterns allegedly made by a computer (i.e., the uncreative ones). These findings are broadly consistent with the intentional or

Overall, the present study experimentally established quality attributions as a causal mechanism that underlies the effects of complex patterns that differ in randomness, in line with our Hypotheses 2-4, thus supporting the aesthetic quality model. The main limitation of the present study is that some of the effects may have been false positive findings, given that not all the *p*-values for the effects we tested were equally convincing (Simonsohn, Nelson, & Simmons, 2014). To address this limitation, in Study 4 we replicated the present findings using a larger sample.

## Study 4

The main aim of the present study was to replicate the results of Study 3 using a larger sample. We found this particularly important because we used priming as a technique to experimentally demonstrate the mechanism, and priming effects have been generally susceptible to replication failures (Cesario, 2014; O’Donnell et al., 2018; Ramscar, 2016). We again predicted that the difference in perceived beauty between the (complex) high and low randomness patterns would be smaller in the incompatible condition compared to either the compatible or the neutral condition. Moreover, concerning the manipulation check, we expected that the difference in quality between these two types of patterns would be smaller in the incompatible than either the compatible or neutral condition.

### Method

#### *Determining Sample Size*

We aimed to conduct a highly powered study ($1 - \beta = .95$) to replicate the effects of Study 3. To this end, we recruited a sample size roughly twice the size of Study 3 (i.e., 500 participants). Based on the data from the previous study, in which 25% participants were

historical theory of art (Levinson, 2002; Bloom, 1996; Bullot & Reber, 2013), according to which an artwork or a visual image (e.g., a random pattern, a non-random pattern) should be perceived as more beautiful if associated with human agency (e.g., when created by humans with the purpose to be creative). However, the findings do not allow distinguishing between the roles that creativity versus human agency play in influencing beauty, and the aesthetic quality model will need to be developed beyond its current state to outline the function that agency plays in the perception of beauty alongside complexity and randomness.

excluded from statistical analyses after the exclusion criteria were applied, we estimated that testing 500 participants would eventually result in roughly 375 participants being included in statistical analyses. We then conducted a sensitivity power analysis using G*Power (Faul et al., 2007) to compute the smallest effect that could be detected using this sample size with the power of .95 and significance criterion of .05. *ANOVA: Fixed effects, omnibus, one-way* was selected, sample size was set to 375, and number of groups to 3. This analysis showed that the study would be sufficiently powered to detect Cohen's *f* equal to 0.204. This effect size is smaller than the effect size that was detected in the previous study (i.e., $\eta_p^2$ = .042, i.e., Cohen's *f* = .209).

### *Participants, Design, Procedure, Stimuli, and Measures*

Five hundred UK nationals completed the study (Female = 316, Male = 184; $M_{\text{age}}$ = 41.858; $SD_{\text{age}}$ = 12.884) via Prolific.co. Payment was £2.15. We again used a 3-level between-subjects design with *compatibility* (incompatible vs. compatible vs. neutral) as the independent variable. After the exclusion criteria were applied, 124 participants who did not correctly answer the check items that were identical as in Study 3 (i.e., the two understanding check items, the general understanding check, the three instructed-response items, and the seriousness check) were excluded from statistical analyses, thus resulting in 376 participants who were included (Female = 237, Male = 139; $M_{\text{age}}$ = 42.388; $SD_{\text{age}}$ = 12.917). Therefore, there were 123 participants in the incompatible condition, 121 in the compatible condition, and 132 in the neutral condition. Experimental procedure, stimuli, and measures were identical to those in Study 3.

## Results

### *Preliminary Analyses*

**Missing Data.** In the present study, no missing cases that would prevent computing the main variables used in statistical analyses were identified.

**Pattern Randomness, Beauty, and Quality.** As in the previous study, we tested whether the low-randomness patterns were perceived as more beautiful and judged as being of higher quality than high-randomness patterns using two repeated measures ANOVAs. The first analysis showed that low randomness patterns ($M$ = 11.724; $SD$ = 2.191) were perceived as more beautiful than the high randomness ones ($M$ = 9.276; $SD$ = 2.191), $F$(1, 375) = 117.285, $p < .001$, $\eta_p^2 = .238$. Likewise, the second analysis showed that low randomness patterns ($M$ = 11.508; $SD$ = 2.393) were judged to be of higher quality than random patterns ($M$ = 9.492; $SD$ = 2.393), $F$(1, 375) = 66.724, $p < .001$, $\eta_p^2 = .151$. Finally, quality was strongly correlated with perceived beauty for either random or non-random patterns, $r$(374) = .703, $p < .001$. The aesthetic quality model was therefore again supported.

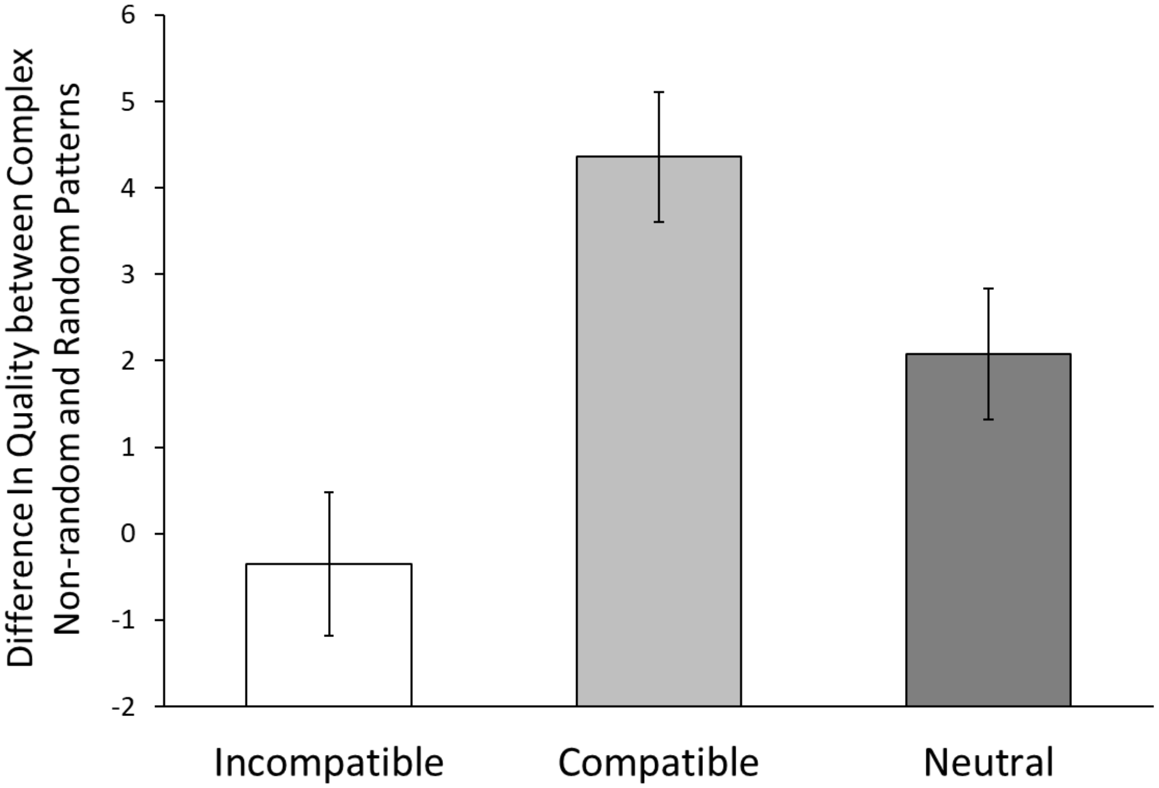


*Figure 14.* Manipulation check: the difference in attributed quality between complex non-random and random patterns as a function of the incompatible, compatible, and neutral

conditions (Study 4). Positive values indicate that participants attributed higher quality to the non-random (relative to random) patterns. Error bars correspond to the 95% Confidence Intervals.

**Manipulation Check.** To test whether *compatibility* condition altered the difference in attributed quality that existed between the low versus high randomness patterns, we performed a one-way ANOVA. The result was highly significant (Figure 14), $F(2, 373) = 34.886$, $p < .001$, $\eta_p^2 = .158$. Planned contrasts further showed that the incompatible condition had a lower difference in quality between complex non-random and random patterns compared to both the compatible, $M_{\text{diff}} = 4.710$, $p < .001$, 95% CI [3.601, 5.819], and neutral condition, $M_{\text{diff}} = 2.431$, $p < .001$, 95% CI [1.346, 3.517], as predicted (Figure 14). The difference between the compatible and neutral conditions for which we did not have a clear prediction was also significant, $M_{\text{diff}} = 2.279$, $p < .001$, 95% CI [1.188, 3.369] (Figure 14). In Supplementary Materials (pp.18-19), we report analyses for creativity and skill manipulation checks individually to show they were impacted by *compatibility* almost identically and were highly correlated, $r(374) = .863$, thus indicating that the manipulations we used tackled quality as a whole.

### *Main Prediction: Compatibility and Differences in Beauty between Patterns*

We ran a one-way ANOVA to test if the difference in beauty between low and high randomness patterns varied across *compatibility* conditions. The results showed a significant effect of compatibility, $F(2, 373) = 13.433$, $p < .001$, $\eta_p^2 = .067$ (Figure 15). Planned contrasts confirmed that the incompatible condition featured a smaller difference in beauty between low and high randomness patterns relative to both the compatible, $M_{\text{diff}} = 2.816$, $p < .001$, 95% CI [1.748, 3.885], and neutral condition, $M_{\text{diff}} = 1.426$, $p = .008$, 95% CI [0.380, 2.472], as predicted (Figure 15). The difference between the compatible and neutral conditions, for which we did make a prediction was also significant, $M_{\text{diff}} = 1.390$, $p = .010$, 95% CI [0.340, 2.441]. Additional analyses probing how specific combinations of compatibility and pattern randomness impacted beauty judgments are available in Supplementary Materials (pp.19-21).

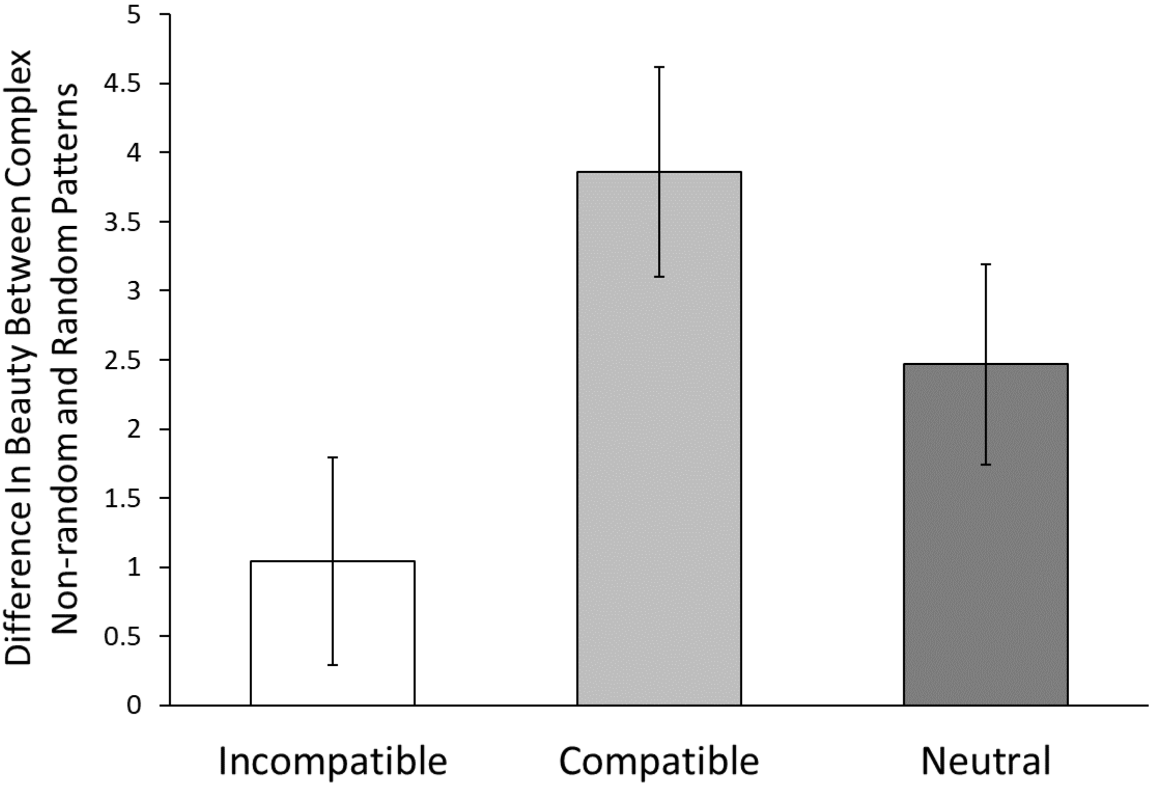


*Figure 15.* The difference in beauty between complex non-random and random patterns as a function of the incompatible, compatible, and neutral conditions (Study 4). Positive values indicate that participants attributed higher beauty to the non-random (relative to random) patterns. Error bars correspond to the 95% Confidence Intervals.

### *Exploratory Analyses*

In exploratory analyses, we examined whether the individual differences measures tested (see the Measures section) would moderate the influence of *compatibility* on the dependent variable. The interaction between each variable and compatibility was computed in a separate multiple regression analysis, and as in Study 3 no interactions were significant, all *p*s ≥ .069. In Supplementary Materials (pp.21-22), we also report additional analyses testing the differences between the patterns we used as stimuli concerning subjective complexity and randomness.

### Discussion

The present study replicated Study 3. As predicted, the incompatible condition decreased the difference in perceived beauty between the complex patterns characterized by low vs. high randomness. Moreover, as expected, the incompatible condition also decreased the difference in attributed quality between the two types of patterns compared to either the compatible or neutral condition. Overall, in line with Hypotheses 2-4 and the aesthetic quality model, the present study established that quality comprises the mechanism that drives the effects on beauty of low versus high randomness among complex patterns.

## General Discussion

The present research provides foundational evidence for complexity and randomness being interactive factors that produce beauty. Specifically, we found that the most beautiful black and white patterns were consistently the ones that were high in complexity but low in randomness. In Study 1, this finding was demonstrated for the patterns from previous research (Chipman, 1977), whereas in Study 2 it was obtained on a more representative population of these patterns, thus showing that our results are not just an artefact created by particular stimuli. We also investigated whether the predictions of our aesthetic quality model would be moderated by individual differences that are typically linked to aesthetic preferences, such as openness to experience (Furnham & Walker, 2001a; Kandler et al., 2016) or political orientation (Furnham & Walker, 2001b; Wilson et al., 1973). However, we did not find convincing evidence that would replicate across studies in support of this possibility. Therefore, the present research indicates that our model of beauty is reasonably generalizable across these individual differences.

The present research also supported our theoretical rationale behind the impact of complexity and randomness on beauty: that complex but non-random patterns are perceived as the most beautiful ones because people associate them with quality (i.e., creativity combined with skill; Kozbelt, 2004). In Study 2, we demonstrated this by showing that perceived quality statistically mediated the impact of the interaction between complexity and randomness on beauty. However, considering various issues associated with mediation analysis (e.g., Bullock, Green, & Ha, 2010; Fiedler, Harris, & Schott, 2018; Fiedler, Schott, & Meiser, 2011), in Studies 3 and 4 we experimentally manipulated the mechanism. In both studies, we showed that, when the originally less beautiful complex random patterns were associated with quality (i.e., participants were told that these patterns were designed to be creative and imaginative), and the originally more beautiful complex non-random patterns were associated with low quality (i.e., participants were told that the patterns were generated to be uncreative and unimaginative), the difference in perceived beauty between the two pattern types decreased. Therefore, our research offers a robust support for the hypothesized mechanism because we demonstrated it in three studies using different methodological procedures.

## Contributions

This research spawns several important theoretical and methodological contributions. On a theoretical level, psychological scientists have striven to identify fundamental visual underpinning of beauty for decades, starting with Berlyne (1963, 1970, 1973, 1974), who proposed one of the first profound theories on complexity as a key quality that predicts beauty. However, empirical tests of his prediction about the inverted-U relationship between complexity and beauty spawned mixed findings (e.g., Silvia, 2005). This led researchers to propose additional constructs that may help explain the link between complexity and beauty, such as order or randomness (e.g., Van Geert & Wagemans, 2020, 2021), but without specifying a clear model revealing how these visual characteristics should combine to predict beauty and why.

To devise a fundamental model that would explain beauty in relation to visual characteristics of an image, we combined literatures on visual complexity, randomness, and quality (e.g., Chipman, 1977; Hagtvedt et al., 2008 Kozbelt, 2004; Van Geert & Wagemans, 2020, 2021; Van Tilburg & Igou, 2014). The model is based on a key assumption that quality (i.e., creativity and skill) determines perceived beauty, and that complexity and randomness interact in predicting beauty because they serve as key indicators of quality. This model goes beyond previous theorizing both because it offers a clear pattern of how complexity and randomness jointly shape beauty, and because it identifies a key mechanism that underpins this influence. Therefore, the present research advances scientific understanding of beauty and links it to basic visual characteristics of a pattern.

The aesthetic quality model was developed with the main aim to explain the link between complexity and beauty. For this reason, we primarily focused on comparing it to other theorizing concerning this link (e.g., Berlyne, 1963, 1970, 1973, 1974; Van Geert & Wagemans, 2020, 2021). However, to appraise our model as a general theory of beauty, it is important to compare it to other theories as well. Perhaps the most influential theoretical account in this regard has been the processing fluency theory (Reber, Schwarz, & Winkielman, 2004), which posits that “the more fluently perceivers can process an object, the more positive their aesthetic response” (p.364). In line with this proposition, visual features such as complexity and creativity should increase beauty only if they evoke fluency, rather than dysfluency (Christensen et al., 2020). Whereas our model, relative to the fluency theory, offers a more nuanced explanation on the link between complexity and beauty, it is plausible that fluency is the final pathway of the aesthetic quality model. For example, quality evoked by non-random but complex patterns may increase beauty because it activates fluent perceptual processing. In that regard, our model may be a more specific case of the fluency theory that focuses on complexity and randomness, and it may explain low-level dimensions of beauty (e.g., valence) that arise during immediate perceptions of stimuli, in line with the pleasure-interest model of aesthetic liking that extends the basic fluency theory (Graf & Landwehr, 2015). Overall, the aesthetic quality model may to some degree overlap with the fluency theory while offering more precise predictions regarding complexity and beauty, and more research will need to be undertaken to integrate the two models.

Concerning methodological contributions, the main one is that, to test our theory, we developed a procedure in Study 2 that can generate a more representative set of binary patterns based on objective indicators of complexity (Chipman, 1977) and randomness. Specifically, rather than creating patterns ourselves or generating an enormous number of random pattens in the hope that some would satisfy the complexity and randomness criteria, we used a process of mutation and selection to generate patterns. In social and cognitive psychology, it is a common problem that stimuli on which researchers test their predictions may not be representative of a general population of these stimuli, which can lead to biased findings (Westfall et al., 2015). To our knowledge, in previous research on beauty, various stimuli were used, from patterns to images (e.g., Forsythe et al., 2011; Newman & Bloom, 2012; Pelowski et al., 2018; Westphal-Fitch & Fitch, 2017). These stimuli were typically created by following a certain rationale or selected from websites or amongst various artworks, but it was not considered whether they exemplify the entire stimuli population to which they belong. It is possible that this could have to some degree accounted for the previously discussed inconsistencies in findings on the link between complexity, randomness, and beauty. Beyond allowing us to test our hypotheses in a way that overcomes this limitation, the procedure we developed can advance research on visual aesthetics more generally by allowing other researchers to test their predictions on representative stimuli sets.

Another important methodological contribution is that we developed a measure of randomness of binary patterns based on Fourier transformations. This was to some degree a necessity because there is a lack of such measures in the literature. Our "Fourier randomness" was a stronger predictor of subjective randomness than another already existing measure we have identified created by Falk and Konold (1997). Therefore, this objective indicator of randomness we created may be considered by other researchers interested in exploring how this quality shapes the perception of beauty and thus advance future research on perception and aesthetics, be it visual or otherwise.

## Limitations and Future Directions

One of the main methodological strengths of the present research is also to some extent its limitation. We argued that using binary patterns to test our theorizing allowed us to precisely compute their randomness and complexity, and to generate a representative population of the stimuli based on different combinations of these two qualities. However, many visual images are not binary, and at present we must be cautious in generalizing our findings to other works of art. In addition, whereas we probed our predictions on black and white binary patterns consisting of 36 squares (6 × 6), such patterns can also have many different sizes and color combinations, which has implications for both their randomness and complexity. Therefore, one possible negative implication of our methodological approach is that the inverted-U relationship (Berlyne, 1963, 1970, 1973, 1974) between complexity and beauty might have failed to occur because the range of stimuli did not include sufficient complexity levels. However, it is important to emphasize that this limitation broadly applies to most papers published in experimental psychology, given that for practical reasons researchers can expose participants to only a range of stimulus values in a set of studies, rather than immediately coming close to exhausting the totality of these stimuli. In line with this premise, rigorously testing whether our hypothesis applies to patterns of various dimensions and color combinations, and then to different visual images beyond patterns, is a long and effortful endeavor that cannot be achieved in a single article, and we see the present research as a starting point of this more elaborate long-term investigation. Here we provide some ideas about how future research could build upon our findings.

The next step of understanding the generalizability of our model could be to repeat the same experiments that we did, but on several other black and white pattern sizes, ranging from 12×12 squares all the way to 1200×1200 squares, which is a typical computer screen resolution and therefore each square would correspond to a pixel. Then, all these experiments could be repeated on different color configurations, starting with only one color in combination with white, and then gradually adding more colors. The final stage of this research endeavor could involve taking real artworks and decomposing them into patterns of squares of the corresponding colors. In these more advanced investigations, there are new challenges that would need to be resolved, such as potentially developing precise measures of complexity that would account for the colors, and calculating how representative specific artworks are of the entire population of stimuli characterized by different combinations of randomness and complexity. Whereas this is a research agenda that could take years, the present research has established a solid starting point that can potentially be taken forward in many ways.

Another potential limitation is that our stimuli had a specific resolution (e.g., 499 × 499 pixels) and were displayed to participants on a computer screen. Given that a digital mode of presentation has become ubiquitous in the current digital age, we do not see our choice to display stimuli on the screen as a limitation but as an ecologically valid methodology. Concerning the stimuli resolution—there are several reasons why this should not be a weakness. First, Chipman (1977) showed that either pattern resolution or context of presentation had little influence of participants' evaluations. Second, the correlation between subjective complexity ratings in our Study 1 and in Chipman (1977) was extremely high (*r* = .891), and the ratings were almost identical despite the studies being around forty years apart and despite the patterns in Chipman (1977) being displayed on the paper and having different resolutions than ours. Therefore, the chance that pattern resolution and mode of presentation could have confounded the findings is minimal.

An additional limitation is that we did not formally incorporate the process of pattern creation into the aesthetic quality model. That is, the model specifies which configurations of randomness and complexity should result in highest beauty (i.e., high complexity and low randomness), but it does not predict that how stimuli are generated should matter, unless this changes their levels of complexity and randomness. However, the manipulations we used in Studies 3-4 to evoke low versus high creativity place the process of creation (i.e., human versus robot) at the basis of the story, and hence questions about the role of this process in the context of our model naturally arise. Moreover, the intentional or historical theories of art (Levinson, 2002; Bloom, 1996; Bullot & Reber, 2013) posit that agency behind artwork creation should impact beauty. Our manipulations are broadly aligned with the notion that processes associated with agency, such as creativity, increase beauty, but we do not further investigate whether separately manipulating parameters such as complexity or creativity and how an image was created (e.g., by a human, machine, natural process, etc.) should change the predictions of our model. This is something that future research and theorizing should address.

A final minor limitation is that we did not test our hypotheses on different cultures, and yet it was shown that Eastern and Western cultures may perceive and evaluate art differently (Masuda, Gonzalez, Kwan, & Nisbett, 2008). Therefore, another step for future research could be to investigate whether the present findings, which were obtained on predominantly Western participant samples, would replicate in an Eastern country such as China or Japan.

## Conclusion

Identifying fundamental visual properties that shape the perception of beauty has captivated humanity for millennia. In the present article, we developed an aesthetic quality model, according to which high complexity combined with low randomness signals quality (i.e., creativity coupled with skill) and leads to an image being perceived as more beautiful. In four studies, we supported this model by using black and white binary patterns as stimuli. This research provides fundamental insights into the perception of beauty and offers several theoretical and methodological advancements that can potentially propel future research on aesthetics.

## References

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management, 39*, 1490-1528.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45*, 527-535.

Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology, 36*, 147-168.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research, 40*, 373-400.

Berlyne, D. E. (1963). Complexity and incongruity variables as determinants of exploratory choice and evaluative ratings. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 17*, 274–290.

Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Perception & Psychophysics*, *8*, 279-286.

Berlyne, D. E. (1973). *Aesthetics and psychobiology*. New York: Appleton Century-Crofts.

Berlyne, D. E. (1974). *Studies in the new experimental aesthetics. Steps toward an objective psychology of aesthetic appreciation*. Washington, DC: Hemisphere.

Bertamini, M., Makin, A., & Rampone, G. (2013). Implicit association of symmetry with positive valence, high arousal and simplicity. *I-Perception, 4*, 317-327.

Birkhoff, G. D. (1932). *Aesthetic Measure*. Cambridge, Mass: Harvard Press.

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition, 60*, 1-29.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what’s the mechanism? (don’t expect an easy answer). *Journal of Personality and Social Psychology, 98*, 550–558.

Bullot, N. J., & Reber, R. (2013). The artful mind meets art history: Toward a psycho-historical framework for the science of art appreciation. *Behavioral and Brain Sciences, 36,* 123-180.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40-48.

Chan, D. W., & Zhao, Y. (2010). The relationship between drawing skill and artistic creativity: Do age and artistic involvement make a difference? *Creativity Research Journal*, *22*, 27-36.

Chipman, S. F. (1977). Complexity and structure in visual patterns. *Journal of Experimental Psychology: General*, *106*(3), 269-301.

Chipman, S. F. (2013, August 2). *Beyond Berlyne's Conjecture: The Aesthetic Quality of Visual Patterns*. Paper presented at Cognitive Science Society, Berlin, Germany.

Chirumbolo, A., Brizi, A., Mastandrea, S., & Mannetti, L. (2014). 'Beauty Is No Quality in Things Themselves': Epistemic Motivation Affects Implicit Preferences for Art. *PLoS one*, *9*.

Christensen, B. T., Ball, L. J., & Reber, R. (2020). Perceptual fluency effects in judgments of creativity and beauty: creative objects are perceived fluently yet they are visually complex. *Journal of Cognitive Psychology*, *32*, 45-66.

Cochran, W. T., Cooley, J. W., Favin, D. L., Helms, H. D., Kaenel, R. A., Lang, W. W., ... & Welch, P. D. (1967). What is the fast Fourier transform? *Proceedings of the IEEE, 55*, 1664-1674.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* NewYork, NY: Routledge Academic.

Di Dio, C., Macaluso, E., & Rizzolatti, G. (2007). The golden beauty: brain response to classical and renaissance sculptures. *PloS one*, *2*, e1201.

Donderi, D. C. (2006). Visual complexity: A review. *Psychological Bulletin*, 132, 73–97.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*, 301-318.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175-191.

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests–An analysis of articles published in 2015. *Journal of Experimental Social Psychology*, *75,* 95-102.

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*, 1231-1236.

Finch, W. H., Bolin, J. E., & Kelley, K. (2019). *Multilevel modeling using R (Second Edition)*. New York: CRC Press.

Forsythe, A., Nadal, M., Sheehy, N., Cela-Conde, C. J., & Sawey, M. (2011). Predicting beauty: fractal dimension and visual complexity in art. *British Journal of Psychology*, *102*, 49-70.

Friedenberg, J. (2019). Beauty in the eye of the beholder: Individual differences in preference for randomized visual patterns. *Experimental Psychology*, *66,* 1-14.

Friedenberg, J., & Liby, B. (2016). Perceived beauty of random texture patterns: A preference for complexity. *Acta Psychologica, 168,* 41-49.

Furnham, A., & Rao, S. (2002). Personality and the aesthetics of composition: A study of Mondrian and Hirst. *North American Journal of Psychology, 4,* 233-242.

Furnham, A., & Walker, J. (2001a). Personality and judgements of abstract, pop art, and representational paintings. *European Journal of Personality*, *15,* 57-72.

Furnham, A., & Walker, J. (2001b). The influence of personality traits, previous experience of art, and demographic variables on artistic preference. *Personality and Individual Differences, 31*, 997-1017.

Gabriel, R. P., & Quillien, J. (2019). A Search for Beauty/A Struggle with Complexity: Christopher Alexander. *Urban Science*, *3*, 64.

Gawronski, B., & Bodenhausen, G. V. (2005). Accessibility effects on implicit social cognition: the role of knowledge activation and retrieval experiences. *Journal of Personality and Social Psychology, 89*, 672-685.

Graf, L. K., & Landwehr, J. R. (2015). A dual-process perspective on fluency-based aesthetics: The pleasure-interest model of aesthetic liking. *Personality and Social Psychology Review*, *19*, 395-410.

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in personality*, *37*, 504-528.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046.

Güçlütürk, Y., Jacobs, R. H., & Lier, R. V. (2016). Liking versus complexity: decomposing the inverted U-curve. *Frontiers in Human Neuroscience*, *10*, 112.

Hager, M., Hagemann, D., Danner, D., & Schankin, A. (2012). Assessing aesthetic appreciation of visual artworks—The construction of the Art Reception Survey (ARS). *Psychology of Aesthetics, Creativity, and the Arts, 6*, 320–333.

Hagtvedt, H., Patrick, V. M., & Hagtvedt, R. (2008). The perception and evaluation of visual art. *Empirical Studies of the Arts*, *26*, 197-218.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making, 8*, 188-201.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, *32*, 385-410.

Hekkert, P., & Van Wieringen, P. C. (1990). Complexity and prototypicality as determinants of the appraisal of cubist paintings. *British Journal of Psychology*, *81*, 483-495.

Hekkert, P., & Van Wieringen, P. C. (1996). Beauty in the eye of expert and nonexpert beholders: A study in the appraisal of art. *The American Journal of Psychology*, 389-407.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid.* Basic Books.

Hübner, R., & Fillinger, M. G. (2016). Comparison of objective measures for predicting perceptual balance and visual aesthetic preference. *Frontiers in Psychology, 7,* 335.

Imamoglu, Ç. (2000). Complexity, liking and familiarity: architecture and non-architecture Turkish students' assessments of traditional and modern house facades. *Journal of Environmental Psychology*, *20*, 5-16.

Jakesch, M., & Leder, H. (2015). The qualitative side of complexity: Testing effects of ambiguity on complexity judgments. *Psychology of Aesthetics, Creativity, and the Arts, 9*, 200–205.

Jander, O. (1991). Rhythmic Symmetry in the" Goldberg Variations". *The Musical Quarterly, 75*, 188-193.

Kandler, C., Riemann, R., Angleitner, A., Spinath, F. M., Borkenau, P., & Penke, L. (2016). The nature of creativity: The roles of genetic factors, personality traits, cognitive abilities, and environmental sources. *Journal of Personality and Social Psychology, 111*, 230–249.

Kant, I. (2000). Critique of the power of judgment. New York, NY: Cambridge University Press.

Kozbelt, A. (2004). Originality and technical skill as components of artistic quality. *Empirical Studies of the Arts*, *22*, 157-170.

Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T. W. (2004). The effort heuristic. *Journal of Experimental Social Psychology*, *40*, 91-98.

Kung, F. Y., Kwok, N., & Brown, D. J. (2018). Are attention check questions a threat to scale validity? *Applied Psychology*, *67*, 264-283.

Kutay, M. A., & Ozaktas, H. M. (1998). Optimal image restoration with the fractional Fourier transform. *JOSA A*, *15,* 825-833.

Lane, S. P., & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships, 35*, 7-31.

Leder, H., & Nadal, M. (2014). Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode–Developments and challenges in empirical aesthetics. *British Journal of Psychology*, *105*, 443-464.

Leder, H., Belke, B., Oeberst, A., & Augustin, M. D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology, 95*, 489-508.

Levinson, J. (2002). The irreducible historicality of the concept of art. *The British Journal of Aesthetics*, *42,* 367-379.

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education, 6,* 8.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1,* 86-92.

Makin, A. D. J., Pecchinenda, A., & Bertamini, M. (2012). Implicit affective evaluation of visual symmetry. *Emotion, 12*, 1021-1030.

Mastandrea, S., Bartoli, G., & Bove, G. (2009). Preferences for ancient and modern art museums: Visitor experiences and personality characteristics. *Psychology of Aesthetics, Creativity, and the Arts, 3*, 164-173.

Masuda, T., Gonzalez, R., Kwan, L., & Nisbett, R. E. (2008). Culture and aesthetic preference: Comparing the attention to context of East Asians and Americans. *Personality and Social Psychology Bulletin*, *34*, 1260-1275.

Mather, G. (2018). Visual image statistics in the history of Western art. *Art & Perception, 8,* 97–115.

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, *97*, 951-966.

Mayer, J., Khairy, K., & Howard, J. (2010). Drawing an elephant with four complex parameters. *American Journal of Physics, 78,* 648-649.

Mayer, S., & Landwehr, J. R. (2018). Quantifying visual aesthetics based on processing fluency theory: Four algorithmic measures for antecedents of aesthetic preferences. *Psychology of Aesthetics, Creativity, and the Arts*, *12*, 399-431.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437-455.

Nadal, M., Munar, E., Marty, G., & Cela-Conde, C. J. (2010). Visual complexity and beauty appreciation: Explaining the divergence of results. *Empirical Studies of the Arts, 28,* 173-191.

Newman, G. E., & Bloom, P. (2012). Art and authenticity: The importance of originals in judgments of value. *Journal of Experimental Psychology: General, 141,* 558–569.

O’Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., ... & Zrubka, M. (2018). Registered replication report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science, 13*, 268-294.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17,* 776-783.

Orne, M. T. (2009). Demand characteristics and the concept of quasi-controls. In R. Rosenthal, & R. L. Rosnow (Eds.). *Artifact in behavioral research* (pp. 143–179). Oxford, UK: Oxford University Press.

Ostrofsky, J., & Shobe, E. (2015). The relationship between need for cognitive closure and the appreciation, understanding, and viewing times of realistic and nonrealistic figurative paintings. *Empirical Studies of the Arts, 33,* 106-113.

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual Review of Psychology, 64,* 77-107.

Pelowski, M., Markey, P. S., Forster, M., Gerger, G., & Leder, H. (2017). Move me, astonish me… delight my eyes and brain: The Vienna integrated model of top-down and bottom-up processes in art perception (VIMAP) and corresponding affective, evaluative, and neurophysiological correlates. *Physics of Life Reviews, 21,* 80-125.

Pelowski, M., Markey, P. S., Goller, J., Förster, E. L., & Leder, H. (2018). But, how can we make “art?” Artistic production versus realistic copying and perceptual advantages of artists. *Psychology of Aesthetics, Creativity, and the Arts*, 462-481.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Van Willigen, B., & Maintainer, R. (2020). Package‘nlme’. Linear and Nonlinear Mixed Effects Models, Version, 3.1–148.

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology, 66*, 29-38.

Ramscar, M. (2016). Learning and the replicability of priming effects. *Current Opinion in Psychology, 12*, 80-84.

Rawlings, D. (2000). The interaction of openness to experience and schizotypy in predicting preference for abstract and violent paintings. *Empirical Studies of the Arts, 18*, 69-91.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8,* 364-382.

Rietzschel, E. F., Nijstad, B. A., & Stroebe, W. (2007). Relative accessibility of domain knowledge and creativity: The effects of knowledge activation on the quantity and originality of generated ideas. *Journal of Experimental Social Psychology, 43,* 933-946.

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences, 50*, 90-94.

Rosenbloom, T. (2006). Color preferences of high and low sensation seekers. *Creativity Research Journal, 18*, 229-235.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12,* 347-367.

Serafin, J., Kozbelt, A., Seidel, A., & Dolese, M. (2011). Dynamic evaluation of high- and low-creativity drawings by artist and nonartist raters: Replication and methodological extension. *Psychology of Aesthetics, Creativity, and the Arts, 5,* 350–359.

Sherman, A., Grabowecky, M., & Suzuki, S. (2015). In the working memory of the beholder: Art appreciation is enhanced when visual complexity is compatible with working memory. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 898-903.

Shariff, A. F., Willard, A. K., Andersen, T., & Norenzayan, A. (2016). Religious priming: A meta-analysis with a focus on prosociality. *Personality and Social Psychology Review*, *20*, 27-48.

Silvia, P. J. (2005). Emotional responses to art: From collation and arousal to cognition and emotion. *Review of General Psychology*, *9*, 342-357.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534-547.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*, 845-851.

Struk, A. A., Carriere, J. S., Cheyne, J. A., & Danckert, J. (2017). A short boredom proneness scale: Development and psychometric properties. *Assessment*, *24*, 346-359.

Swami, V., & Furnham, A. (2012). The effects of symmetry and personality on aesthetic preferences. *Imagination, Cognition and Personality*, *32*, 41-57.

Teigen, K. H. (1994). Yerkes-Dodson: A law for all seasons. *Theory & Psychology*, *4*, 525-547.

Tempelaars, S. (1996). *Signal processing, speech and music*. Lisse, Netherlands: Swets & Zeitlinger Publishers.

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184-197.

Tinio, P. P., & Leder, H. (2009). Just how stable are stable aesthetic features? Symmetry, complexity, and the jaws of massive familiarization. *Acta Psychologica*, *130*, 241-250.

Van Geert, E., & Wagemans, J. (2020). Order, complexity, and aesthetic appreciation. *Psychology of Aesthetics, Creativity, and the Arts, 14*, 135-154.

Van Geert, E., & Wagemans, J. (2021). Order, complexity, and aesthetic preferences for neatly organized compositions. *Psychology of Aesthetics, Creativity, and the Arts, 15*, 484–504.

Van Koningsbruggen, G. M., Stroebe, W., & Aarts, H. (2011). Through the eyes of dieters: Biased size perception of food following tempting food primes. *Journal of Experimental Social Psychology, 47*, 293-299.

Van Tilburg, W. A. P., & Igou, E. R. (2014). From Van Gogh to Lady Gaga: Artist eccentricity increases perceived artistic skill and art appreciation. *European Journal of Social Psychology, 44*, 93-103.

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science, 10*, 390-399.

Westphal-Fitch, G., & Fitch, W. T. (2017). Beauty for the eye of the beholder: Plane pattern perception and production. *Psychology of Aesthetics, Creativity, and the Arts, 11*, 451–456.

Wicks, R. (1995). Kant on fine art: Artistic sublimity shaped by beauty. *The Journal of Aesthetics and Art Criticism, 53*, 189-193.

Wiersema, D. V., Van Der Schalk, J., & van Kleef, G. A. (2012). Who's afraid of red, yellow, and blue? Need for cognitive closure predicts aesthetic preferences. *Psychology of Aesthetics, Creativity, and the Arts, 6*, 168-174.

Wilson, G. D., Ausman, J., & Mathews, T. R. (1973). Conservatism and art preferences. *Journal of Personality and Social Psychology, 25*, 286-288.

Winograd, S. (1978). On computing the discrete Fourier transform. *Mathematics of Computation, 32*, 175-199.

Wundt, W. (1874). *Grundzüge der physiologischen psychologie* [Fundamentals of physiological psychology]. Leipzig, Germany: Wilhelm Engelmann.