The Nature-Disorder Paradox: A Perceptual Study on How Nature is Disorderly Yet Aesthetically Preferred

Hiroki P. Kotabe, Omid Kardan, & Marc G. Berman

University of Chicago

Author Note

All authors, Department of Psychology, University of Chicago, 5848 South University Avenue, Chicago, IL 60637.

This work was supported by a grant from the TKF Foundation to M.G.B., two grants from the John Templeton Foundation (the University of Chicago Center for Practical Wisdom and the Virtue, Happiness, and Meaning of Life Scholars Group) to M.G.B., and a grant from the National Science Foundation (BCS-1632445) to M.G.B., as well as an internal grant from the University of Chicago to M.G.B.

Thank you to Daniel Casasanto, Reid Hastie, and Wilhelm Hofmann for their invaluable support as members of H.P.K.'s dissertation committee. Thank you to the editor and the reviewers for their excellent feedback. And thank you to Daniel Hayes and Lillian Li for their research assistance.

An earlier version of this paper was included as a chapter of H.P.K.'s dissertation.

Correspondence and requests for materials should be addressed to H.P.K. (hkotabe@uchicago.edu) or M.G.B. (bermanm@uchicago.edu).

Abstract

Natural environments have powerful aesthetic appeal linked to their capacity for psychological restoration. In contrast, disorderly environments are aesthetically aversive, and have various detrimental psychological effects. But in our research, we have repeatedly found that *natural environments are perceptually disorderly*. What could explain this paradox? We present three competing hypotheses: The aesthetic preference for naturalness is more powerful than the aesthetic aversion to disorder (the *nature-trumps-disorder hypothesis*); disorder is trivial to aesthetic preference in natural contexts (the *harmless-disorder hypothesis*); and disorder is aesthetically preferred in natural contexts (the *beneficial-disorder hypothesis*). Utilizing novel methods of perceptual study and diverse stimuli, we rule in the nature-trumps-disorder hypothesis and rule out the harmless-disorder and beneficial-disorder hypotheses. In examining perceptual mechanisms, we find evidence that high-level scene semantics are both *necessary* and *sufficient* for the nature-trumps-disorder effect. Necessity is evidenced by the effect disappearing in experiments utilizing only low-level visual stimuli (i.e., where scene semantics have been removed) and experiments utilizing a rapid-scene-presentation procedure that obscures scene semantics. Sufficiency is evidenced by the effect reappearing in experiments utilizing noun stimuli which remove low-level visual features. Furthermore, we present evidence that the interaction of scene semantics with low-level visual features *amplifies* the nature-trumps-disorder effect—the effect is weaker both when statistically adjusting for quantified low-level visual features and when using noun stimuli which remove low-level visual features. These results have implications for psychological theories bearing on the joint influence of low- and high-level perceptual inputs on affect and cognition, as well as for aesthetic design.

*Keywords***:** naturalness**,** disorder, aesthetics, scene aesthetics, visual perception

> Nature holds the key to our aesthetic, intellectual, cognitive and even spiritual satisfaction.
>
> —E. O. Wilson

There are multifold benefits of exposure to natural environments (Berman, Jonides, & Kaplan, 2008; Berman et al., 2012; Berto, 2005; Bratman, Daily, Levy, & Gross, 2015; Cimprich & Ronis, 2003; Kaplan & Berman, 2010; Kardan, Gozdyra, et al., 2015; Kuo & Sullivan, 2001a, 2001b; Ulrich, 1984), whereas exposure to disorderly environments has a variety of detrimental effects (Chae & Zhu, 2014; Geis & Ross, 1998; Heintzelman, Trent, & King, 2013; Keizer, Lindenberg, & Steg, 2008; Kotabe, 2014; Perkins & Taylor, 1996; Ross, 2000; Tullett, Kay, & Inzlicht, 2015; Vohs, Redden, & Rahinel, 2013; Wilson & Kelling, 1982). Exposure to natural environments may improve health (Kardan, Gozdyra, et al., 2015), increase physical activity (Humpel, Owen, & Leslie, 2002), improve memory and attention (Berman et al., 2008; Berman et al., 2012), boost positive affect (Berman et al., 2012), alleviate negative affect (Bratman et al., 2015), and decrease aggression and crime (Kuo & Sullivan, 2001a, 2001b). On the contrary, exposure to disorderly environments, may diminish a sense of meaning in life (Heintzelman et al., 2013), elicit negative affect (Ross, 2000; Tullett et al., 2015), reduce self-control and cognitive-control (Chae & Zhu, 2014), and encourage rule-breaking and criminal behavior (Keizer et al., 2008; Kotabe, Kardan, & Berman, 2016).

Nature's restorative potential has been theoretically and empirically linked with a strong aesthetic preference for natural environments (Han, 2010; Hartig & Staats, 2006; Purcell, Peron, & Berto, 2001; Staats, Van Gemerden, & Hartig, 2010; Ulrich, 1983; Van den Berg, Koole, & van der Wulp, 2003). 'Aesthetic preference' refers to a 'like-dislike' affective response (Zajonc, 1980) elicited by visual exposure to scenes (Ulrich, 1983). It may be separable from other components of reward such as 'wanting' and 'learning' (Berridge, Robinson, & Aldridge, 2009). Scores of studies suggest that natural environments tend to be aesthetically preferred over built environments (for reviews, see R. Kaplan & Kaplan, 1989; Ulrich, 1983), and aesthetic preference for natural environments over built environments is so strong that often distributions for aesthetic preference ratings between these two environmental categories hardly overlap (S. Kaplan, Kaplan, & Wendt, 1972; Ulrich, 1983). In contrast, research on visual aesthetics suggests that disorderly environments are aesthetically aversive because of their lack of spatial structure (Palmer, Schloss, & Sammartino, 2013) and because of the disfluent experience of viewing them (Arnheim, 1974; Kotabe, Kardan, & Berman, 2016b; Reber, Schwarz, & Winkielman, 2004).

But, paradoxically, *nature is perceived as disorderly*. We have found repeatedly comparing naturalness and disorder judgments for large and diverse sets of scene images that naturalness and disorder are significantly correlated (correlations ranging from .35 to .42). How is it that nature scenes have strong aesthetic appeal when they are perceptually disorderly? One possibility is that the positive effect of naturalness on aesthetic preference trumps the negative effect of disorder (the *nature-trumps-disorder hypothesis*). That is, aesthetic preference for naturalness and aesthetic aversion to disorder may operate more or less independently, but aesthetic preference for nature is more powerful than aesthetic aversion to disorder, thus natural scenes can be disorderly yet aesthetically preferred. Natural scenes may in part have powerful aesthetic appeal because of 'biophilia'—a powerful affinity to the natural and the living that is rooted in our evolutionary history (Wilson, 1984). According to this hypothesis, natural scenes would have powerful aesthetic appeal because of their association with life and survival. Largely left out of the discussion, however, is the role of the basic physical or low-level visual features of the environment (Berman et al., 2012; Berman, Jonides, & Kaplan, 2008; S. Kaplan, 1995). In this study, ‘low-level visual features’ refers collectively to the basic spatial and color features of a scene (e.g., edges, hue). Low-level visual features are involved in the early stages of perceiving semantic features.

Low-level visual features are important not only for aesthetic preference for natural scenes (Kardan et al., 2015) but also for the perception of naturalness itself (Berman et al., 2014; see also Torralba & Oliva, 2003). Berman and colleagues (2014) showed that naturalness was related to the density of contrast changes (i.e., straight and non-straight edges) in the scene, the average color saturation of the scene, and the hue diversity of the scene. A machine-learning classification algorithm based on these features could predict whether an image was perceived as natural or built with 81% accuracy. Of particular interest is that the strongest low-level visual predictor of naturalness was the density of non-straight edges which include curved contours. Research suggests that people prefer curved contours to sharp contours because the latter are threatening (Bar & Neta, 2006, 2007), and thus, the abundance of curved contours and relative absence of sharp contours in nature may be important to nature’s aesthetic potency. Furthermore, Kardan and colleagues (2015) showed that naturalness modeled by these low-level visual features could predict aesthetic preference. To be clear, some of the relationship between low-level visual features and aesthetic preference may be mediated by higher-level scene semantics (e.g., vegetation, water, sky), but there are direct effects to varying degrees as well (Ibarra et al., 2017), which may be due to some low-level visual features being imbued with meaning themselves (Kotabe, Kardan, & Berman, 2016a; see also Bar & Neta, 2006, 2007). All of this research points to the possibility that nature’s beauty may not be entirely about biophilic responses to high-level scene semantics, but also responses to low-level visual features. It is unclear, however, whether the low-level visual features embedded in nature scenes alone can drive a strong aesthetic preference through their associations with naturalness, or rather if the interaction with scene semantics is of particular importance. It may be that low-level visual features *amplify* the effect of naturalness on aesthetic preference. That is, compared to sensory perceptions, mental representations may be more like “cardboard cutouts of reality” (Gilbert & Wilson, 2007; cf. Kosslyn, 1996) and thus scene semantics of nature scenes on their own may not have quite the impact on aesthetic preference as real nature scenes, which possess rich spatial, color, and semantic features.

In addition to the nature-trumps-disorder hypothesis, there are two plausible alternative explanations for the nature-disorder paradox. First, disorder may have a negligible effect on aesthetic preference in natural environments (the *harmless-disorder hypothesis*). That is, disorder may be aesthetically aversive in built environments but trivial to aesthetic preference in natural environments (see R. Kaplan & Austin, 2004). This could be due to people *expecting* natural environments to be disorderly and responding neutrally to the status quo (e.g., a typical unstructured nature scene). In contrast, if people expect built environments to be orderly, they may respond negatively when that expectation is disconfirmed (e.g., when seeing a dilapidated building). There is abundant evidence for such confirmation bias and belief perseverance (Nickerson, 1998), and the assumption here is that these tendencies plays a role in the formation of aesthetic preference for scenes. Second, disorder may actually be aesthetically *preferred* in natural environments (the *beneficial-disorder hypothesis*). That is, disorder may be aesthetically aversive in built environments but aesthetically preferred in natural environments, thus nature scenes could be aesthetically preferred in part because they are disorderly (see Özgüner & Kendle, 2006; Van den Berg & van Winsum-Westra, 2010). This could be due to disorderly and "wild" nature being reminiscent of ancestral environments that helped sustain human life (e.g., densely vegetated areas providing food and shelter) (E. O. Wilson, 1984; see also Appleton, 1996).

Testing these three competing hypotheses and examining at what level of visual perception they operate would not only help us make sense of the nature-disorder paradox but would generally be informative to psychological theories concerning the joint influence of lower- and higher-level perceptual inputs on affect and cognition. Little work has systematically separated the low- and high-level inputs of environmental scenes, much less tested whether there are differential effects of distinct low- vs. high-level inputs vs. their interactions on important everyday psychological experiences such as like-dislike affective responses. Many insights can be gleaned from examining such questions. For example, if low-level visual features and perceived disorder contribute to aesthetic preferences for nature scenes, it would provide further evidence against the idea that natural environments compose a monolithic category and have uniform effects on people (see Ulrich, 1983). If the effect of naturalness on aesthetic preference depends on the level of perceptual input, or the interaction between levels of perceptual input, it would support our position that naturalness and its aesthetics are complex and nuanced, dependent on the interplay of lower- and higher-level perceptual inputs (Berman et al., 2014; Ibarra et al., 2017; Kardan et al., 2015). Simply finding that disorder affects aesthetic preferences for natural scenes would answer the important yet unanswered question, does disorder matter in nature? Surprisingly little is known about this because virtually all of the research on environmental disorder has sampled stimuli only from the domain of built environments.

In the following series of experiments, we tested the three competing hypothesis using diverse stimuli and novel methods of perceptual study and found converging evidence supporting the nature-trumps-disorder hypothesis and disconfirming the harmless-disorder and beneficial-disorder hypotheses. Furthermore, we show that when scene semantics are obscured, the nature-trumps-disorder effect does not hold, whereas when low-level visual features are obscured, the nature-trumps-disorder effect is preserved but attenuated. These results suggest that scene semantics are *necessary and sufficient* for the nature-trumps-disorder effect, and that low-level visual features *amplify* this effect when interacting with scene semantics.

General Method

We sampled broadly from real-world environments by using diverse sets of images of environmental scenes (see Figure 1 for examples; all images utilized in this study can be downloaded here in original resolution: goo.gl/za9seG).[1] One set contained 260 scene images ranging from more built to more natural according to previously collected ratings (see Berman et al., 2014; Kardan et al., 2015). Another set contained 916 images selected from the Scene UNderstanding (SUN) image database (http://vision.princeton.edu/projects/2010/SUN/) (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010) that were even more diverse in semantic content (e.g., nature-related scene images contained not only trees, parks, etc. but also waves, mountains, and lava). In our first experiments, we took a principled scene-statistics approach (Geisler, 2008) to analyzing the basic physical properties of these scenes to shed light on various questions bearing on the nature-disorder paradox such as the validity of the competing hypotheses and the extent to which low- vs. high-level perceptual mechanisms are at work. First, in Experiments 1a-c and Experiments 2a-c, we used the scene images in their unaltered form and quantified their low-level visual features to statistically estimate contributions of low-level visual features and high-level scene semantics to aesthetic preferences. Next, in Experiments 3a-f, we scrambled low-

[1] Regarding the ecological validity of scene images, it was shown that walking in natural vs. urban environments has similar effects on directed attention as viewing images of natural vs. urban environments (Berman et al., 2008).

level spatial (Experiments 3a-c) and color (Experiments 3d-f) features from the scene images to assess the effects of obscuring scene semantics. In Experiments 4a-c, we took an alternative approach to addressing the effects of obscuring scene semantics by rapidly presenting scene images which can obscure scene semantics while preserving all of the basic physical properties of the scene images. Across all of these experiments, we had people rate naturalness, disorder, and aesthetic preference for the given stimulus set-type. For these experiments, data analysis was conducted at the image-level. Lastly, in Experiments 5a-c, we conducted a similar set of experiments except that noun stimuli were used instead of images to assess the effect of obscuring low-level visual features while preserving scene semantics. For these experiments, data analysis was conducted at the word-level.

[INSERT FIGURE 1 HERE]

Quantifying ‘Naturalness’ and ‘Disorder’

By ‘naturalness’ and ‘disorder’ we are referring to *subjective judgments* about a scene or derived stimuli at the level of a global description. In our research, we have found that when a person is presented a scene image, they can quickly and spontaneously form judgments about its level of ‘naturalness’ and ‘disorder’. We have collected many thousands of such ratings without directing participants with explicit definitions for these dimensions. By analyzing these spontaneous ratings in relation to low-level visual features of the scenes and several semantic judgments, we have found clear systematicity in ratings from diverse participants across diverse scenes, and thus have been able to make progress toward quantitative definitions of naturalness (Berman et al., 2014) and disorder (Kotabe et al., 2016b).

Regarding the quantification of naturalness, Berman et al. (2014) utilized both computational (machine learning) and explicit-rating approaches to quantify basic spatial and color features of hundreds of scene images and used these features to predict naturalness judgments for those scenes. First, they implicitly defined naturalness using a multidimensional scaling analysis of people's spontaneous arrangements of the similarity of scenes, and found that the primary dimension was related to naturalness according to free-labeling of this dimension from an independent set of raters. Second, they explicitly defined naturalness ratings by first having people rate the scene images on a naturalness scale and then predicted whether a scene was perceived as natural based on quantified low-level visual features such as edge density, color saturation, and hue diversity.

Regarding the quantification of disorder, Kotabe et al. (2016b) also took an explicit rating approach in which they predicted disorder ratings for hundreds of scene images using quantified low-level visual features. The features that best predicted disorder judgments were non-straight edge density and reflectional asymmetry. To estimate the reliability of the edge features in determining perceived disorder, a series of experiments were conducted in which these features were extracted and scrambled and participants rated the resulting stimuli in terms of disorder. Even though participants could not make out the scenes from which the edge features originated, their disorder ratings for these low-level visual features predicted the disorder ratings of the original unaltered scenes. Furthermore, a new set of stimuli was created based on manipulating non-straight edge density and asymmetry and these were rated in terms of disorder by an independent set of raters. These two low-level spatial features had large and predictable effects on disorder ratings.

We note that there are numerous ways to quantify the spatial and color properties of scene images, and thus, there are likely additional low-level visual predictors of naturalness and disorder judgments that have not yet been identified. Regarding spatial features, our decision to focus on edge features and (a)symmetry was guided by our goal to analyze features that are easily translatable to design applications. Other spatial features such as holistic textural properties proposed by Oliva and Torralba (2001) have various uses, such as for computer vision, but would be more difficult to translate for design purposes. Regarding color features, much less work has been done on the visual perception of color features, therefore we defaulted to using color features based on the standard hue-saturation-value (HSV) model of the RGB color space.

Experiments 1a-c: Reanalyzing Previously Collected Data

As a first test of the nature-disorder paradox and the three competing hypotheses, we reanalyzed previously collected naturalness, disorder, and aesthetic preference ratings for 260 environmental scene images (naturalness and aesthetic preference ratings from Kardan et al., 2015; disorder ratings from Kotabe et al., 2016b). We also quantified spatial and color visual features as in Berman et al. (2014) and Kardan et al. (2015). By statistically adjusting for low-level visual variation in the environmental scene images, we could conduct an initial test of the extent to which the relative effects of naturalness and disorder on aesthetic preference depend on low-level visual features. This would shed light on whether nature's aesthetic appeal indeed depends not only on high-level scene semantics but also low-level visual features as suggested by prior work from our lab (Berman et al., 2014; Ibarra et al., 2017; Kardan et al., 2015).

Scene Selection

The scene images utilized in this work were the same as in Berman et al. (2014) and Kardan et al. (2015). The selection criteria for these images targeted diversification on the naturalness dimension, which was validated in the aforementioned study by Berman et al. (2014). The images depicted scenes from Nova Scotia, urban parks from Annapolis, Baltimore, and Washington D.C., and various everyday settings in Ann Arbor, Detroit, and Chicago. Only scenes without humans or animals present were selected.

Scene Ratings

Naturalness, disorder, and aesthetic preference were all assessed with seven-point bipolar scales. The naturalness scale was anchored with endpoints labeled “very manmade” and “very natural”, the disorder scale was anchored with endpoints labeled “very orderly” and “very disorderly”, and the aesthetic preference scale was anchored with endpoints labeled “strongly dislike” and “strongly like”. Simple like-dislike ratings of this kind reliably reflect affective discriminations (Zajonc, 1980). Naturalness and aesthetic preference ratings were collected in a physical laboratory setting (see Kardan et al., 2015 for full procedural details). Scene images were presented in full resolution (512*384, 685*465, or 1024*680 pixels) on a plain white background for 1 s and then removed from the screen. Participants were then given up to 4 s to make a rating for each scene. Each participant rated all 260 scene images in random order with naturalness ratings and aesthetic preference ratings made in counterbalanced blocks. Disorder ratings were collected in an online experiment (see Kotabe et al., 2016b for full procedural details). In this experiment, each participant rated a random subset of 50 of the 260 scene images (10 randomly selected from each quintile of previously collected naturalness ratings) presented in random order. Scene images were presented in a 600*450 pixel frame on a plain white background and participants had unlimited time to rate each image. The rating scale was presented below the image and participants could make a rating at any time. Because differences in stimulus size and duration were a potential issue, we conducted Experiments 2a-c in which we conceptually replicated Experiments 1a-c using a new and larger set of scene images presented with identical stimulus sizes and durations across different rating tasks.

Quantifying Low-Level Visual Features

We utilized MATLAB's Image Processing Toolbox to quantify four low-level spatial features and six low-level color features of the scene images. The spatial features quantified were non-straight edge density (a measure of how many non-straight edges are in the scene image), straight edge density (a measure of how many straight edges are in the scene image), vertical reflectional asymmetry ("vertical asymmetry" for short; a measure of how well the left and right halves of the scene image mirror each other), and horizontal reflectional symmetry ("horizontal symmetry" for short; a measure of how well the top and bottom halves of the scene image mirror each other). Both faint and salient edge features were detected using the Canny edge detection algorithm (Canny, 1986) and straight edges were quantified with a connected components algorithm based on the extent to which an edge's coordinates varied perpendicular to its direction (see Kardan et al., 2015). The resulting color features, based on the standard Hue-Saturation-Value (HSV) model, were mean hue (a measure of the average color appearance of a scene), mean saturation (a measure of how intense or pure the colors of the scene are on average), and mean value (a measure of the average luminance of a scene), as well as the standard deviations of those color measures as measures of hue diversity, saturation diversity, and value diversity. Straight edge density, non-straight edge density, saturation, value, SD saturation, and SD value were all quantified from their respective maps created as in Berman et al. (2014) and Kardan et al. (2015). Because the hue of a pixel is an angular value, mean and SD hue were calculated using circular statistics (Circular Statistics Toolbox for MATLAB, Berens, 2009). Asymmetry was quantified by summing up the dot product of the left and mirrored-right half (vertical symmetry) or the top and mirrored-bottom half (horizontal symmetry) of the edge map of the scene images. These sums were then normalized to a [0 1] range by being divided by the total number of non-zero pixels in the edge map of the corresponding image (i.e., the total edge space).

Results and Discussion

First, we examined correlations to test for the nature-disorder paradox. Naturalness and disorder were significantly correlated at $r = .35$, $p < .001$ (see Table 1 for correlation matrices of naturalness, disorder, and aesthetic preference ratings across all experiments; see supplementary materials for descriptive statistics and scatterplots of these ratings across all experiments). Naturalness was significantly correlated with aesthetic preference, $r = .73$, $p < .001$ but disorder was not significantly correlated with aesthetic preference, $r = -.08$, $p = .177$. After statistically adjusting for disorder, naturalness was partially correlated with aesthetic preference, $r_p = .81$, $p < .001$ and, after statistically adjusting for naturalness, disorder was partially correlated with aesthetic preference ratings, $r_p = -.52$, $p < .001$. The positive correlation between naturalness and disorder and the contradirectional correlations with preference demonstrate the nature-disorder paradox.

Table 1
Correlations Between Naturalness, Disorder, and Aesthetic Preference Ratings Across all Experiments

	Experiments 1a-c (260 scenes)			Experiments 2a-c (916 scenes)		
	Naturalness	Disorder	Aesthetic Preference	Naturalness	Disorder	Aesthetic Preference
Naturalness	–			–		
Disorder	.35***	–		.36***	–	
Aesthetic	.73***	-.08	–	.46***	-.16***	–

Preference						
	Experiments 3a-c (260 scrambled-edge stimuli)			Experiments 3d-f (260 scrambled-color stimuli)		
Naturalness	–			–		
Disorder	NA	–		-.31***	–	
Aesthetic Preference	NA	-.64***	–	.02	-.36***	–
	Experiments 4a-c (260 inverted scenes, 50 ms)			Experiments 5a-c (632 nouns)		
Naturalness	–			–		
Disorder	-.17	–		.37***	–	
Noun Preference	.04	-.07	–	.34***	-.22***	–
*** p < 0.001						

Note. "NAs" for correlations with naturalness ratings in Experiments 3a-c (260 scrambled-edge stimuli) because of low rater consistency. *** *p* < .001

Next, we tested the three competing hypotheses. In order to compare the relative importance of concepts measured on different scales for aesthetic preference, we simultaneously regressed aesthetic preference on naturalness, disorder, and their interaction and tested the relative importance of each factor by comparing standardized coefficients (see Table 2, Experiments 1a-c, Model 1). These factors explained almost two thirds of the variance in aesthetic preference, $R^2_{\text{adj}} = .65$. Both naturalness, $\beta = 0.88$, $t(256) = 21.61$, $p < .001$, $\eta_p^2 = .65$, and disorder, $\beta = -0.39$, $t(256) = -9.93$, $p < .001$, $\eta_p^2 = .28$, significantly predicted aesthetic preference. A linear contrast indicated that the effect of naturalness on aesthetic preference was significantly larger than the effect of disorder, $F(1, 256) = 117.17$, $p < .001$, supporting the nature-trumps-disorder hypothesis. The relative importance of naturalness and disorder for predicting aesthetic preference was estimated with the *relaimpo* R package (Grömping, 2006), which implements eight methods of estimating relative importance that take into account intercorrelations between explanatory variables. Across all eight metrics, naturalness was estimated to be more important than disorder in terms of explaining aesthetic preference—e.g., the popular *lmg* method (Lindeman, Merenda, & Gold, 1980), which partitions R^2 by averaging over orders, estimated that 90% of the variance in the aesthetic preference model was explained by naturalness ratings vs. 10% by disorder ratings (see Table 3 to compare with other experiments). Regarding the harmless-disorder and beneficial-disorder hypotheses, both of these hypotheses would predict a positive interactive effect between naturalness and disorder on aesthetic preference ratings. There was actually a marginal *negative* interaction between naturalness ratings and disorder ratings, $\beta = -0.08$, $t(256) = -1.85$, $p = .066$, $\eta_p^2 = .01$, suggesting, if anything, that disorder may have a slightly stronger negative effect on aesthetic preference in natural environments than in built environments—contrary to the harmless-disorder and beneficial-disorder hypotheses.

Table 2

Regression Models, Experiments 1a-c and 2a-c

		Experiments 1a-c (260 scenes)		Experiments 2a-c (916 scenes)	
		Model 1 (R^2_{adj} = .65)	Model 2 (R^2_{adj} = .70)	Model 1 (R^2_{adj} = .33)	Model 2 (R^2_{adj} = .44)
High-level scene semantics	Naturalness	0.88*** (0.04)	0.84*** (0.05)	0.60*** (0.03)	0.58*** (0.04)
	Disorder	-0.39*** (0.04)	-0.39*** (0.04)	-0.37*** (0.03)	-0.40*** (0.03)
	Nature × disorder interaction	-0.08^ (0.04)	-0.09* (0.04)	0.03 (0.03)	0.02 (0.03)
Low-level spatial features	Non-straight edge density		0.11 (0.08)		0.19* (0.09)
	Straight-edge density		0.05 (0.05)		0.04 (0.05)
	Vertical symmetry		0.05 (0.06)		-0.13* (0.06)
	Horizontal symmetry		0.18** (0.06)		0.13* (0.05)
Low-level color features	Hue		0.03 (0.04)		-0.01 (0.03)
	Saturation		0.14** (0.05)		0.13*** (0.04)
	Value		0.01 (0.04)		-0.10** (0.03)
	SD hue		0.16** (0.05)		-0.00 (0.03)
	SD saturation		0.05 (0.05)		0.04 (0.03)
	SD value		-0.05 (0.04)		0.10** (0.03)
*** p < 0.001, ** p < 0.01, * p < 0.05 ^ p < .10					

Note. Standardized coefficients not in parentheses and standard errors in parentheses. * p < .05,** p < .01, *** p < .001

Table 3

Relative Importance Estimates of Naturalness and Disorder for Aesthetic Preference When Scene Semantics are Salient (Experiments 1a-c, 2a-c, and 5a-c)

	Adjusting for low-level visual features	Naturalness	Disorder	Difference
Experiments 1a-c (260 scenes)	No	90%	10%	80%
	Yes	63%	9%	54%
Experiments 2a-c (916 scenes)	No	77%	23%	54%
	Yes	41%	20%	21%
Experiments 5a-c (632 nouns)	NA	58%	37%	21%

Note. Positive difference indicates the nature-trumps-disorder effect. Remarkably, the difference score in Experiments 2a-c (916 scenes) when adjusting for low-level visual features was virtually equal to the difference score in Experiments 5a-c in which we used 632 noun stimuli, providing converging evidence for the validity of these approaches for estimating the effect of obscuring low-level visual features. Furthermore, the 26% reduction in difference score due to adjusting for low-level visual features in Experiments 1a-c (260 scenes) is similar to the 33% reduction in difference score due to adjusting for low-level visual features in Experiments 2a-c (916 scenes), suggesting that low-level visual features amplified the nature-trumps-disorder effect to a similar degree between these two sets of experiments which used different scene images, different procedures, and different participant samples.

To estimate the independent effects of high-level naturalness and disorder scene semantics, we statistically adjusted for the quantified spatial and color low-level visual features by including these features as predictors in another multiple regression model (R^2_{adj} = .70, see Table 2, Experiments 1a-c, Model 2). Both naturalness, β = 0.84, $t(246) = 16.11$, $p < .001$, $\eta_p^2 =$ .51, and disorder, β = -0.39, $t(246) = -9.72$, $p < .001$, $\eta_p^2 = .28$, still significantly predicted aesthetic preference, and a linear contrast again indicated that the effect of naturalness on aesthetic preference was significantly larger than the effect of perceived disorder, $F(1, 246) =$ 63.85, $p < .001$. Regarding relative importance, the *lmg* method estimated that 63% of the variance in the aesthetic preference model was explained by naturalness vs. 9% by disorder. Furthermore, there was still a small but significant negative interaction between naturalness and disorder, $\beta = -0.09$, $t(246) = -2.12$, $p = .035$, $\eta_p^2 = .02$, contrary to the harmless-disorder and beneficial-disorder hypotheses.

It is noteworthy that adjusting for low-level visual features decreased the explanatory power of naturalness in predicting aesthetic preference from 90% to 63% but only decreased the explanatory power of disorder in predicting aesthetic preference from 10% to 9%. This result suggests that low-level visual features play an asymmetric role in the relationships between naturalness and aesthetic preference vs. the relationship between disorder and aesthetic preference—with low-level visual features playing a larger role in naturalness predicting aesthetic preference than in disorder predicting aesthetic preference. Although nature scene semantics have a larger effect on aesthetic preference, the low-level visual features embedded in natural scenes seem to make an important contribution. That said, there are some methodological issues with reanalyzing these data, which warrant reservations, which we resolve in the following conceptual replication.

Experiments 2a-c: Conceptual Replication

We resolved issues with reanalyzing data in the previous set of experiments by conducting a conceptual replication in Experiments 2a-c. First, the selection criteria for the scene images used in Experiments 1a-c targeted diversification on the naturalness dimension rather than both on this dimension and on the disorder dimension. Such sampling bias could cause external validity issues (Brunswik, 1947; Wells & Windschitl, 1999), though we note that these are correlated dimensions and thus sampling on one dimension samples on the other. Experiments 2a-c further address this issue by using a larger and more diverse sample of scene images selected based on criteria targeting diversification on both the naturalness and disorder dimensions. Second, the image rating task differed on some procedural parameters (e.g., stimulus duration, stimulus size) between Experiments 1a-c, so it was important to ensure that these differences were not confounding the results by using the same image rating task parameters across different rating tasks in Experiments 2a-c. Third, in Experiments 1a-c, aesthetic preference and naturalness ratings were collected from a different population (i.e., college students) from the disorder ratings (i.e., online sample more representative of the U.S. population), which is resolved in Experiments 2a-c by sampling participants from the same population.

Method

Participants and design. 702 US-based adults (392 women, 308 men, 2 other) were recruited from the online labor market Amazon Mechanical Turk (AMT) and were randomly assigned to one of the three sub-experiments—rating naturalness (Experiment 2a), disorder (Experiment 2b), or aesthetic preference (Experiment 2c). Sample size and stopping rule were based on our goal to receive ~20 ratings per image. Ages ranged from 18 to 76 (*M* = 36.39, *SD* = 12.73). 555 participants identified primarily as White/Caucasian, 54 as Black/African American, 39 as Asian/Asian American, 37 as Hispanic/Latino, 8 as “multiple ethnicities,” 5 as Native American/Alaska Native, and 3 as “other.” Participants were compensated $1.00 for their participation and the experiment lasted for approximately 20 minutes. Informed consent was administered by the Institutional Review Board (IRB) of the University of Chicago.

Scene selection. Scene images were selected from the SUN image database (Xiao et al., 2010); a database that contains a more semantically diverse set of images than was used in Experiments 1a-c (e.g., including scenes of open sky, waves, and volcanoes). Selection criteria targeted diversification on both the naturalness and disorder dimensions, with an emphasis on increasing the representation of orderly nature scenes and disorderly built scenes as compared with the set of 260 scenes used previously. As in Experiments 1a-c, only scenes without humans or animals were selected. This yielded a set of 1,105 scene images which included orderly and disorderly nature scenes as well as orderly and disorderly built scenes.

Procedure. Participants were first given a brief introduction to the image-rating task. They were then presented a randomly selected 100 of the 1105 scene images in a 720*540 pixel frame on a plain white background. The given rating scale was positioned immediately below each scene image. Participants were given unlimited time to make each rating. As in the reanalyzed disorder-rating experiment, we decided not to use time restrictions across rating tasks to capture participants' spontaneous assessments of naturalness, disorder, and aesthetic preference. We again did not provide any explicit definition of naturalness or disorder because our goal here was to test for systematicity in people's *spontaneous* perceptions of disorder and naturalness.

Regarding the rating scales, we closely followed the previously used procedure. In the naturalness experiment (Experiment 2a), participants were asked, "How manmade or natural does this environment look to you?" In the disorder experiment (Experiment 2b), participants were asked, "How disorderly or orderly does this environment look to you?" And in the aesthetic preference experiment (Experiment 2c), participants were asked, "How much do you dislike or like this environment?" Participants made ratings using seven-point scales ("very manmade" to "very natural"; "very disorderly" to "very orderly"; "strongly dislike" to "strongly like"). In addition, an independent sample of participants did a fourth version of this experiment in which they rated "rule-breaking" which is a complex concept beyond the scope of this study, because here we focus on physical disorder rather than social forms of disorder which may have little to do with the basic physical features of the scene. Thus, we strictly limited the presence of rule-breaking by only including images which rated less than 2 on the 1-7 rule-breaking scale ("no rule-breaking" to "a lot of rule-breaking"), leaving 916 images for our statistical analysis.

Results and Discussion

Because participants were sampled from a diverse online sample and rated different scene images (due to randomly presenting a subset of the scene images to each participant), it was important to test rater consistency. Rater consistency was estimated with Shrout and Fleiss's (1979) Case 2 intraclass correlation formula for average measures which utilizes a two-way random effects model in which image and rater are both modeled as random effects. For naturalness ratings, the consistency estimate was *ICC* = .99, 95% CI [.99, .99]; for disorder ratings, the consistency estimate was *ICC* = .95, 95% CI [.95, .96]; and for aesthetic preference ratings, the consistency estimate was *ICC* = .94, 95% CI [.94, .95], all of which would be considered high reliability estimates by conventional standards (Cicchetti, 1994).

First, we tested for the nature-disorder paradox. Naturalness and disorder were again significantly correlated, $r = .36$, $p < .001$ (see Table 1). The degree of correlation between naturalness and disorder in this set of experiments was remarkably close to the degree of correlation between naturalness and disorder observed in Experiments 1a-c ($r = .35$), even when this set of experiments was not a direct replication but rather a conceptual replication using different scene images, different procedures, and different participant samples, attesting to the robustness of the relationship between naturalness and disorder. Naturalness was significantly correlated with aesthetic preference, $r = .46$, < .001, and disorder was significantly correlated with aesthetic preference at $r = -.16$, $p < .001$. After adjusting for disorder, naturalness was partially correlated with aesthetic preference at $r_p = .56$, $p < .001$ and, after adjusting for naturalness, disorder was partially correlated with aesthetic preference at $r_p = -.40$, $p < .001$. The positive correlation between naturalness and disorder and the contradirectional correlations with preference again demonstrate the nature-disorder paradox.

Next, we tested the three competing hypotheses. As before, we simultaneously regressed aesthetic preference ratings on naturalness ratings, disorder ratings, and their interaction (see Table 2, Experiments 2a-c, Model 1). These factors explained about a third of the variance in aesthetic preference ratings, $R^2_{\text{adj}} = .33$. This is about half of the variance in aesthetic preference explained in Experiments 1a-c, likely because we sampled much more diverse scene images. Both naturalness ratings, $\beta = .60$, $t(912) = 20.47$, $p < .001$, $\eta_p^2 = .32$, and disorder ratings, $\beta = -.37$, $t(912) = -12.57$, $p < .001$, $\eta_p^2 = .15$, again significantly predicted aesthetic preference ratings. A linear contrast indicated that the effect of naturalness on aesthetic preference was significantly larger than the effect of perceived disorder, $F(1, 912) = 43.01$, $p < .001$, again supporting the nature-trumps-disorder hypothesis. We estimated the relative importance of naturalness and disorder for predicting aesthetic preference as before. Across all eight metrics calculated by the *relaimpo* package, naturalness was estimated to be more important than disorder for aesthetic preference—e.g., the *lmg* method estimated that 77% of the variance in the model was explained by naturalness vs. 23% by disorder. The 54% difference score estimates the size of the nature-trumps-disorder effect. Regarding the alternative hypotheses, there was no significant interaction between naturalness and disorder, $\beta = .03$, $t(912) = 0.96$, $p = .339$, $\eta_p^2 = .00$, providing no support for the harmless-disorder and beneficial-disorder hypotheses.

Adjusting for the quantified low-level visual features in another multiple regression model, both naturalness ratings, $\beta = .58$, $t(902) = 25.83$, $p < .001$, $\eta_p^2 = .22$, and disorder ratings, $\beta = -.40$, $t(902) = -13.71$, $p < .001$, $\eta_p^2 = .17$, still significantly predicted aesthetic preference ratings (see Table 2, Experiments 2a-c, Model 2). Furthermore, a linear contrast indicated that the effect of naturalness on aesthetic preference was still significantly larger than the effect of disorder, $F(1, 902) = 18.57$, $p < .001$, though to a lesser extent than in the previous multiple regression model. Regarding relative importance, the *lmg* method estimated that 41% of the variance in the model was explained by naturalness vs. 20% by disorder. Statistically adjusting for low-level visual features again decreased the explanatory power of naturalness more than disorder—the variance in aesthetic preference explained by naturalness dropped from 77% to 41% whereas the variance in aesthetic preference explained by disorder dropped only from 23% to 20%, again suggesting an asymmetric role of low-level visual features in naturalness vs. disorder in predicting aesthetic preference. Regarding the harmless-disorder and beneficial-disorder hypotheses, there was no significant interaction between naturalness ratings and disorder ratings, $\beta = .02$, $t(902) = 0.72$, $p = .473$, $\eta_p^2 = .00$, providing no support for these hypotheses.

The size of the nature-trumps-disorder effect can be estimated by taking the difference between the relative importance estimates of naturalness and disorder for aesthetic preference (see Table 3). In Experiments 2a-c, the nature-trumps-disorder effect size decreased from 54% to 21% after adjusting for low-level visual features. In Experiments 1a-c, the nature-trumps-disorder effect size estimate decreased from 80% to 54% after adjusting for low-level visual features. The absolute change due to statistically adjusting for low-level visual features between these sets of experiments was remarkably similar at 26% in Experiments 1a-c and 33% in Experiments 2a-c, suggesting that low-level visual features amplified the nature-trumps-disorder effect to a similar degree between these two sets of experiments. In Experiments 2a-c (but not in the less-controlled Experiments 1a-c), non-straight edge density significantly predicted aesthetic preference but straight-edge density did not, consistent with work referenced in the introduction regarding the preference for curved contours over sharp contours (Bar & Neta, 2006, 2007). This result suggests that part of aesthetic preference for nature may be due to the presence of curved contours, among other low-level visual features.

Overall, this was a remarkably successful conceptual replication, considering that we used a completely different and more diverse set of scene images, changed the procedural parameters across all of the rating tasks, and sampled participants from a different population for the naturalness and aesthetic preference rating tasks. This conceptual replication lends credence to the idea that nature's powerful aesthetic appeal is a function of both scene semantics and low-level visual features. What is unclear still is whether the contribution of low-level visual features embedded in nature scenes to nature's aesthetic appeal is due to an interaction between these low-level visual features and scene semantics, or if these low-level visual features on their own have a marked effect on aesthetic preference through their association with naturalness. That is, does the nature-trumps-disorder effect hold at the level of low-level visual features when high-level scene semantics are obscured? Answering this question would tell us whether scene semantics are necessary for the nature-trumps-disorder effect. Conversely, does the nature-trumps-disorder effect hold at the level of high-level scene semantics when low-level visual features are obscured? Answering this question would tell us whether scene semantics are sufficient for the nature-trumps-disorder effect. Or, is the interaction between low-level visual features and high-level scene semantics important for the nature-trumps-disorder effect? We addressed these questions in the following series of experiments in which we tested for the nature-trumps-disorder effect under conditions in which scene semantics are obscured via (a) extraction and scrambling of low-level visual features (Experiments 3a-f), (b) scene semantics are obscured via rapid presentation of inverted scenes (Experiments 4a-f), and (c) low-level visual features are obscured via use of noun stimuli (Experiments 5a-c).

Experiments 3a-c: Obscuring Scene Semantics by Extracting and Scrambling Edges

In Experiments 3a-c, we again followed the scene statistics approach (Geisler, 2008) by constructing new stimuli which were derived from the set of 260 scene images by extracting and scrambling only the quantified edge features of those scenes. We had people rate the edge features alone in terms of naturalness (Experiment 3a), disorder (Experiment 3b), or aesthetic preference (Experiment 3c). With these data, we could test whether the nature-trumps-disorder effect holds at the level of edges, when scene semantics are obscured.

Method

Participants and design. 287 US-based adults (159 men, 126 women, 2 other) were recruited from AMT and were randomly assigned to one of the three sub-experiments. Sample size and stopping rule were based on our goal to receive ~20 ratings per image. Ages ranged from 18 to 70 (*M* = 31.71, *SD* = 10.21). 223 participants identified primarily as White/Caucasian, 25 as Asian/Asian American, 19 as Black/African American, 12 as Hispanic/Latino, 6 as “other,” 1 as Native American/Alaska Native, and 1 as Native Hawaiian/Pacific Islander. Participants were compensated $0.50 for their participation and the experiment took approximately 10 minutes. Informed consent was administered by the IRB of the University of Chicago.

Constructing scrambled-edge stimuli. For the scrambled-edge stimuli, we devised a novel method to remove scene semantics while preserving edge formations from the original scene images (see supplementary materials for an illustration of the processes involved in this method). First, we created an edge map from the original scene images, created as in Berman et al. (2014) and Kardan et al. (2015). Next, the edge map of the target image was randomly rotated either 90 or 270 degrees and overlaid on the 180-degrees-rotated edge map, constructing a stimulus comprised of twice as many edges (but the same straight and non-straight edge ratios) as the scene image. A mask matrix was then constructed to be the same size as the scene images (600*800) with its elements randomly assigned between zero and one. This matrix was then convolved with a median filter sized 30*40 pixels. In this way, patches of 1s and 0s were made randomly and placed at random locations across the mask with random sizes equal to or greater than 30*40 pixels, with every mask having, on average, half a surface of 1s and half a surface of 0s. This mask was then multiplied (dot product) by the doubled edge map so that half of its edges were removed at random. The resulting stimulus had, on average, the same amount of edges with similar edge types as the original scene image from which it was derived, but the scene semantics were largely obscured. See Figure 2 (middle panels) for examples.

Procedure. Participants rated the derived scrambled-edge stimuli in terms of naturalness, disorder, and aesthetic preference following the same procedure as in Experiments 2a-c except that scene images were presented in a 600*450 pixel frame.

[INSERT FIGURE 2 HERE]

Results and Discussion

We estimated rater consistency as in Experiments 2a-c. For naturalness ratings, the consistency estimate was *ICC* = .26, 95% CI [.13, .37]; for disorder ratings, the consistency estimate was *ICC* = .90, 95% CI [.88, .92]; and for aesthetic preference ratings, the consistency estimate was *ICC* = .69, 95% CI [.63, .74]. The confidence interval indicates that the consistency estimate for naturalness ratings was significantly below what is conventionally considered “fair” reliability (.40 to .59, Cicchetti, 1994), though a positive estimate suggests some systematicity in these ratings. Although edge features can reliably predict scene naturalness, in isolation, they seem to have a weak naturalness signal (Kotabe et al., 2016a), perhaps because they have a minimal direct effect on perceived naturalness, rather operating through high-level scene semantics (Ibarra et al., 2017). Because of the weak naturalness signal, we could not test for the nature-disorder paradox or the nature-trumps-disorder effect. We know that disorder inversely correlated with aesthetic preference, *r* = -.64, < .001, but for naturalness, the signal was too weak to test the relationship with disorder or aesthetic preference. That said, the weak naturalness signal precludes the presence of a nature-trumps-disorder effect, thus answering the question that motivated this set of experiments, which was whether the nature-trumps-disorder effect would hold at the level of edges, when scene semantics are obscured. Next, we tested whether the nature-trumps-disorder effect holds at the color-level, when scene semantics are obscured.

Experiments 3d-f: Obscuring Scene Semantics by Scrambling Colors

In Experiments 3d-f, we first constructed color stimuli that obscure scene semantics by scrambling the color features of the 260 scene images. We then had participants rate the color features alone in terms of naturalness (Experiment 3d), disorder (Experiment 3e), and aesthetic preference (Experiment 3f). With these ratings we could test whether the nature-trumps-disorder effect holds at the color-level, when scene semantics are obscured.

Method

Participants and design. 288 US-based adults (168 men, 119 women, 1 other) were recruited from AMT and were randomly assigned to one of the three experiments. Sample size and stopping rule were based on our goal to receive ~20 ratings per image. Ages ranged from 18 to 75 (*M* = 32.92, *SD* = 11.20). 223 participants identified primarily as White/Caucasian, 27 as Asian/Asian American, 21 as Black/African American, 11 as Hispanic/Latino, 4 as “other,” and

1 as Native American/Alaska Native. Participants were compensated $0.50 for their participation and the experiment took approximately 10 minutes. Informed consent was administered by the IRB of the University of Chicago.

Constructing scrambled-color stimuli. Constructing the scrambled-color stimuli was a simpler task than constructing the scrambled-edge stimuli. It also did not require as much alteration to the original scene images. To construct the scrambled-color stimuli, we randomly repositioned windows of 5*5 pixels from the scene image. Thus, all pixels from the original scene images were preserved. The window size was selected so that (a) scene semantics would become non-discernable, and (b) the color textures of the scene would be preserved. For example, pretesting revealed that a 1*1 pixel window size resulted in stimuli in which less frequent colors were so scattered that they became invisible to the eye whereas using a 10*10 pixel window kept some of the objects or segments of the scene identifiable. See Figure 2 (right panels) for examples.

Procedure. The procedure was identical to that of Experiments 3a-c except that participants were presented the scrambled-color stimuli instead of the scrambled-edge stimuli.

Results and Discussion

We estimated rater consistency as in the previous experiments. For naturalness ratings, the consistency estimate was *ICC* = .80, 95% CI [.77, .84]; for disorder ratings, the consistency estimate was *ICC* = .66, 95% CI [.60, .71]; and for aesthetic preference ratings, the consistency estimate was *ICC* = .62, 95% CI [.55, .68]. The estimates indicate good to excellent reliability across all ratings (Cicchetti, 1994). The higher consistency estimate for naturalness ratings for scrambled-color stimuli than for scrambled-edge stimuli suggests that naturalness is better preserved in color features than in edge features, consistent with our prior work (Kotabe et al., 2016a).

With all three rating types receiving reliable ratings, we again tested for the nature-disorder paradox. Contrary to the nature-disorder paradox, naturalness ratings and disorder ratings for these stimuli were *inversely* correlated at $r = -.31$, $p < .001$ (see Table 1), suggesting that the color features embedded in natural scenes are associated with *order*. This is an intriguing and paradoxical result in and of itself that requires further research. It raises the question, how are natural scenes disorderly when their color features are orderly? Furthermore, naturalness was *not* significantly correlated with aesthetic preference ratings, $r = .02$, $p = .750$, but disorder ratings were at $r = -.36$, $p < .001$. After adjusting for disorder ratings, naturalness ratings were still not significantly correlated with aesthetic preference ratings, $r = -.10$, $p = .104$, and, after adjusting for naturalness ratings, disorder ratings were partially correlated with aesthetic preference ratings at virtually the same level as before, $r_p = -.37$, $p < .001$. The absence of contradirectional effects of naturalness and disorder on aesthetic preference is inconsistent with the nature-disorder paradox. The absence of the nature-disorder paradox precludes the nature-trumps-disorder effect.

A possible concern with Experiments 3a-c (scrambled-edges) and 3d-f (scrambled-colors) is that the nature-trumps-disorder effect was eliminated not due to obscuring scene semantics, but rather it could be an artifact of substantially altering the original scene images through our novel methods of low-level visual feature extraction. Although we preserved all of the pixels from the scene image in our method of scrambling colors, the resulting stimuli are quite different from the original scene images. Making substantial alterations is necessary to create visual stimuli that largely obscure scene semantics, however, one can also obscure scene semantics by rapidly presenting unaltered scene images below specific presentation times at which certain scene semantics become perceivable (Fei-Fei, Iyer, Koch, & Perona, 2007). To further test whether the nature-disorder paradox and the nature-trumps-disorder effect are eliminated when scene semantics are obscured, we conducted another set of experiments following this alternative procedure.

Experiments 4a-c: Obscuring Scene Semantics via Rapid Presentation of Inverted Scenes

We obscured scene semantics in this set of experiments by rapidly presenting inverted but unaltered scene images for 50 ms. Participants rated naturalness, disorder, and aesthetic preference after rapid exposure to each scene image. Decisions to use 50 ms and scene inversion were largely guided by research by Fei-Fei et al. (2007). In this research, Fei-Fei and colleagues examined what people perceive in a glance at a scene image. Examining natural scenes and objects specifically, they found that after 53 ms scene exposure, peoples' reports of what they saw mostly reflected sensory features of the scenes rather than semantic features such as distinct objects, though there was still some accurate recall of such scene semantics. Therefore, to further obscure scene semantics, we inverted the scene images because rotating familiar objects (e.g., trees, buildings) to unfamiliar orientations makes them more difficult to recognize (Logothetis & Sheinberg, 1996; Yin, 1969). Under these conditions, we could test whether the nature-trumps-disorder effect holds when scene semantics are largely obscured, without altering the original scene images.

Method

Participants and design. 333 US-based adults (193 men, 133 women, 1 other) were recruited from AMT and were randomly assigned to one of the three experiments. Sample size and stopping rule were based on our goal to receive ~20 ratings per image. Ages ranged from 19 to 79 (*M* = 35.85, *SD* = 11.64). 238 participants identified primarily as White/Caucasian, 42 as Asian/Asian American, 22 as Black/African American, 20 as Hispanic/Latino, 2 as Native American, 3 as "multiple," and 3 as "other". Participants were compensated $0.70 for their participation and the experiment took a median of 6 minutes to complete. Informed consent was administered by the IRB of the University of Chicago.

Materials. 260 scene images from Experiments 1a-c rotated 180 degrees. All scene images were preloaded at the beginning of the study while participants read the consent form to prevent delayed presentation during the rapid-scene-presentation task.

Procedure. See Figure 3 for an illustration of a single trial of the rapid-scene-presentation procedure. In a single trial of this task, we presented a fixation cross for 1 s, then the inverted scene image for 50 ms, then a perceptual mask for 1 s, and then a seven-point rating scale to assess naturalness ("very manmade" to "very natural", Experiment 4a), disorder ("very disorderly" to "very orderly", Experiment 4b), or aesthetic preference ("strongly dislike" to "strongly like", Experiment 4c). Participants were given unlimited time to make their ratings. The next trial started automatically after a rating was made. The scene image was masked with one of eight perceptual masks constructed by convolving a random matrix of elements assigned between zero and one with a median filter sized 40*30 pixels (an intermediate step in the edge extraction and scrambling process). The scene images and perceptual masks were presented in 720*540 pixel frames on a white background. Each participant was presented 50 inverted scene images randomly selected from the set of 260 inverted scene images and presented in random order.

[INSERT FIGURE 3 HERE]

Results and Discussion

Before conducting our statistical analysis, we examined presentation times to ensure that the flipped scene images were presented for the targeted amount of time (50 ms), in case the execution of the JavaScript function we wrote for rapidly displaying and then hiding the images erred on occasion. We accurately measured presentation time by taking the difference between the recorded system times at image presentation and image hiding. Per Fei-Fei (2007), we excluded trials in which the presentation time exceeded 53 ms because scene semantics become significantly more recalled and low-level visual features become significantly less recalled at longer exposures. In total, 5.75% of the trials were excluded.

We estimated rater consistency as in the previous experiments. For naturalness ratings, the consistency estimate was *ICC* = .35, 95% CI [.24, .45]; for disorder ratings, the consistency estimate was *ICC* = .37, 95% CI [.26, .46]; and for aesthetic preference ratings, the consistency estimate was *ICC* = .37, 95% CI [.26, .47]. The confidence intervals across these estimates indicate that reliability was not significantly below the conventionally fair range (.40 to .59; Cicchetti, 1994). We note that in this set of experiments we did not expect high consistency across raters because the results of Fei-Fei et al. (2007) suggest that we largely obscured scene semantics already by presenting scene images for only 50 ms, and furthermore, by inverting the scene images. By largely obscuring scene semantics, it follows that we substantially reduced the signal strength of naturalness and disorder, as evidenced by the rater consistency estimates.

We first tested for the nature-disorder paradox. Contrary to the nature-disorder paradox, there was again a significant *inverse* correlation between naturalness and disorder, *r* = -.17, *p* = .006 (see Table 1), mirroring the results from Experiments 3d-f in which we used scrambled-color stimuli. We speculate that when inverted scene images are presented for 50 ms, color features are perceived more than edge features, as detection of edges, by definition, first requires processing discontinuities in color features. Furthermore, naturalness was *not* significantly correlated with aesthetic preference, $r = .04$, $p = .548$, and neither was disorder, $r = -.07$, $p = .234$. After statistically adjusting for disorder, naturalness was still not significantly correlated with aesthetic preference, $r = .03$, $p = .686$, and, after adjusting for naturalness, disorder was still not significantly correlated with aesthetic preference, $r_p = -.07$, $p < .001$. Again, as in Experiments 3d-f, we did not observe contradirectional effects of naturalness and disorder on aesthetic preference indicative of the nature-disorder paradox. As in Experiments 3d-f, the absence of the nature-disorder paradox precludes the nature-trumps-disorder effect.

Using completely different methods, Experiments 3a-c, 3d-f, and 4a-c converge on the finding that the nature-trumps-disorder effect does not hold when scene semantics are largely obscured. In fact, when scene semantics are obscured, the nature-disorder paradox disappears. In Experiments 3a-c (scrambled edges), naturalness signal was reduced to an extent that suggests that the nature-trumps-disorder effect did not hold. In Experiments 3d-f (scrambled colors) and Experiments 4a-c (rapid presentation of inverted scenes), naturalness and disorder were *inversely* correlated, precluding the nature-trumps-disorder effect. Furthermore, in a set of unreported experiments, we followed the same procedure as in Experiments 4a-c, except that we presented scene images in original orientation (not inverted) and for 67 ms. According to Fei-Fei et al. (2007), significantly more scene semantics should be perceived under these conditions, and this is just what we observed. Rater consistency analysis indicated that the naturalness signal was stronger than when we presented inverted scene images for 50 ms and we observed a strong nature-trumps-disorder effect similar to in Experiments 1a-c and 2a-c. Taking into account the evidence presented so far, we conclude that high-level scene semantics are *necessary* for the nature-trumps-disorder effect. But are they also *sufficient*?

Experiments 5a-c: At the Level of Scene Semantics

To test whether scene semantics are sufficient for the nature-trumps-disorder effect, we used noun stimuli instead of scene images. By using noun stimuli, we could convey the semantic features of scenes absent of low-level visual features (they could only be imagined, not perceived, see our note below). In this way, this set of experiments is the counterpart to Experiments 3a-f (low-level visual stimuli) and Experiments 4a-c (rapid presentation of inverted scenes) in which we obscured high-level scene semantics. We presented people with a wide variety of nouns ranging from conveying more natural semantics (e.g., ‘mountain’, ‘tree’, ‘swamp’) to more urban semantics (e.g., ‘office’, ‘factory’, ‘traffic’). Participants rated these nouns either in terms of naturalness (Experiment 5a), disorder (Experiment 5b), or aesthetic preference (Experiment 5c). With these ratings, we could test whether the nature-trumps-disorder effect holds when low-level visual features are obscured.

We note that when forming judgments about nouns, participants may have a mental image of an exemplar of the referent (Paivio, 1969), and thus, low-level visual features may be imagined but cannot be perceived. Neural and behavioral evidence points to some overlap between imagery and visual perception (Kosslyn, 1996), however, that overlap seem to be less pronounced in visual cortex (Ganis, Thompson, & Kosslyn, 2004; Mellet, Tzourio, Denis, & Mazoyer, 1995). Furthermore, we know that by using noun stimuli, participants could not be exposed to low-level visual features per our definition of them as overt physical features of environmental scenes. Our goal here was not to eliminate mental imagery, but rather to test for the presence of the nature-disorder paradox and evaluate the competing hypotheses under conditions which obscure low-level visual features of environmental scenes.

Method

Participants and design. 1,572 US-based adults (861 women, 707 men, 4 other) were recruited from AMT and were randomly assigned to one of the three experiments. Sample size and stopping rule were based on our goal to receive ~100 ratings per noun. Ages ranged from 18 to 85 (*M* = 35.79, *SD* = 13.00). 1,217 participants identified primarily as White/Caucasian, 122 as Black/African American, 96 as Asian/Asian American, 79 as Hispanic/Latino, 41 as "multiple," 10 as Native American, 6 as "other," and 1 as Native Hawaiian. Participants were compensated $0.50 for their participation and the experiment took approximately 10 minutes to complete. Informed consent was administered by the IRB of the University of Chicago.

Materials. In total, 632 nouns were selected from the MRC Psycholinguistic Database (Coltheart, 1981) (see the supplementary materials for a full list of the nouns). Selection criteria targeted diversification on the naturalness dimension.

Procedure. The 632 nouns were split into ten quantiles based on their Thorndike-Lorge written frequency (TL-FRQ) measure (Thorndike & Lorge, 1944). The 10 quantiles of nouns were each placed in a block which also included one attention check item (e.g., "select strongly like so we know you are paying attention."). The purpose of the attention check was to maintain engagement in case rating nouns was less engaging than rating scene images. Participants were randomly presented 10 nouns (or 9 nouns and an attention check item) from each randomly presented quantile, thus each participant could rate 81 to 100 nouns that ranged from less to more common. Nouns were presented on the center of the screen in Arial font, sized 64 pixels. Participants rated naturalness (Experiment 5a), disorder (Experiment 5b), or preference

(Experiment 5c) as in the previous experiments. Also, participants had unlimited time to make each rating as in the previous experiments. The procedures thus closely followed the image rating task procedure except with noun stimuli instead of scene images.

Results and Discussion

We estimated rater consistency as in the previous experiments. For naturalness ratings, the consistency estimate was *ICC* = .99, 95% CI [.99, .99]; for disorder ratings, the consistency estimate was *ICC* = .96, 95% CI [.96, .97]; and for aesthetic preference ratings, the consistency estimate was *ICC* = .95, 95% CI [.94, .95]. As with the scene images, the estimates indicate high reliability for all of these ratings, suggesting that naturalness and disorder once again had strong signal strength.

First, we tested for the nature-disorder paradox. Naturalness and disorder were significantly correlated, $r = .37$, $p < .001$. This correlation was remarkably close to the correlations we observed in Experiments 1a-c ($r = .35$) and Experiments 2a-c ($r = .36$), in which we used two different sets of scene images. Naturalness was significantly correlated with noun preference ratings, $r = .34$, $p < .001$, and disorder was significantly correlated with noun preference ratings, $r = -.22$, $p < .001$. After adjusting for disorder, naturalness was partially correlated with noun preference ratings, $r_p = .46$, $p < .001$, and after adjusting for naturalness, disorder ratings were partially correlated with noun preference ratings, $r_p = -.39$, $p < .001$. The positive correlation between naturalness and disorder and the contradirectional correlations with preference indicating the return of the nature-disorder paradox.

With the nature-disorder paradox present again, we next tested the three competing hypotheses. Noun preference ratings were simultaneously regressed on naturalness ratings, disorder ratings, and their interaction. We statistically adjusted for two factors for which we had data for all of the nouns by including these factors in the regression model (TL-FRQ and word length). This model explained over a quarter of the variance in noun preference ratings, R^2_{adj} = .29. Both naturalness ratings, β = 0.50, t(625) = 13.43, p < .001, η_p^2 = .23, and disorder ratings, β = -0.42, t(625) = -11.52, p < .001, η_p^2 = .17, significantly predicted noun preference ratings. A linear contrast indicated that the effect of naturalness on noun preference was significantly larger than the effect of disorder, F(1, 625) = 4.42, p = .036, indicating the return of the nature-trumps-disorder effect. We also estimated the relative importance of naturalness and disorder for explaining noun preference as before. Across all eight metrics calculated, naturalness was estimated to be more important than disorder for noun preference ratings—e.g., the *lmg* method estimated that 58% of the variance in the preference model was explained by naturalness ratings vs. 37% by disorder ratings. Regarding the harmless-disorder and beneficial-disorder hypotheses, there was a small but significant *negative* interaction between the effects of naturalness and disorder on noun preference, β = -0.19, t(625) = -5.29, p < .001, η_p^2 = .04, mirroring the small negative interaction we observed in Experiments 1a-c, and contradicting the harmless-disorder and beneficial-disorder hypotheses.

Overall, these results are similar to the results of the experiments in which we used scene images as stimuli, except for one important difference. The difference between the relative importance of naturalness vs. disorder for noun preference (58% vs. 37%, respectively; 21% absolute difference) was not nearly as large as the difference between the relative importance of naturalness vs. disorder we observed when we regressed aesthetic preference on naturalness, disorder, and their interaction in Experiments 1a-c (90% vs. 10%, respectively; 80% difference) and in Experiments 2a-c (77% vs. 23%, respectively; 54% difference) in which participants rated scene images (see Table 3 to compare with other experiments). However, when adjusting for low-level visual features in those experiments, the estimated relative importance of naturalness and disorder for aesthetic preference in Experiments 1a-c (63% vs. 9%, respectively; 54% difference) and Experiments 2a-c (41% vs. 20%, respectively; 21% difference) shifted closer to what we observed in the present experiments in which we used noun stimuli. In fact, the difference in relative importance estimates between naturalness and disorder in Experiments 2a-c (916 scene images) when adjusting for low-level visual features was virtually identical to the difference in relative importance estimates between naturalness and disorder in this set of experiments (632 noun stimuli). We conclude that scene semantics seem to be *sufficient* for the nature-trumps-disorder effect. However, scene semantics are not all that matter—the low-level visual features embedded in nature scenes *amplify* the effect.

General Discussion

How are nature scenes disorderly yet aesthetically preferred? In our study, we delved into this question utilizing diverse stimuli and methods of perceptual study. The results of our experiments support the nature-trumps-disorder hypothesis and provide contradictory evidence against the harmless-disorder and beneficial-disorder hypotheses. That is, nature scenes can be disorderly yet aesthetically preferred because the effect of naturalness on aesthetic preference is stronger than the effect of disorder on aesthetic preference, and not because disorder does not matter for nature scenes or because disorder is aesthetically pleasing in nature scenes. Furthermore, the results suggest that nature's full aesthetic appeal depends on the joint influence of scene semantics and low-level visual features, though scene semantics are necessary and sufficient to get the effect. Influential hypotheses such as biophilia (E. O. Wilson, 1984) have emphasized high-level semantic associations with life and survival, while the role of the low-level visual features of the environment has received less attention. Consistent with previous research which suggests an important role of low-level visual features for perceived naturalness (e.g., Berman et al., 2014; Ruderman & Bialek, 1994; Torralba & Oliva, 2003) and for nature's aesthetics (Kardan et al., 2015), we find that the nature-trumps-disorder effect is strongest when both scene semantics and low-level visual features are at play (Experiments 1a-c and 2a-c). In contrast, the nature-trumps-disorder effect is absent when scene semantics are obscured (Experiments 3a-f and 4a-c), and present but attenuated when low-level visual features are obscured (Experiments 5a-c). In summary, we conclude that scene semantics are *necessary and sufficient* for the nature-trumps-disorder effect, and low-level visual features *amplify* the effect.

To our knowledge, this is the first psychological study of the joint influence of naturalness and disorder on aesthetic preferences. Previous psychological research has focused solely on aesthetic preference for natural scenes and environments (Kaplan, Kaplan, & Wendt, 1972; Kardan, Demiralp, et al., 2015; Ulrich, 1983; Van den Berg et al., 2003). The results of this study suggest that it may be fruitful to pursue research at the intersection of these two dimensions, which have been treated in isolation. For example, if disorder has a negative impact on affective responses in natural environments as suggested by this study, it opens up the possibility that there are other psychological and behavioral consequences of disorder in natural environments. The separability of the effects of naturalness and disorder on aesthetic preference suggests that there could be other separable psychological effects operating in parallel. Thus, there could be other puzzling and paradoxical psychological effects of naturalness and disorder. For example, a certain natural environment may be restorative (Berman et al., 2008; Bratman, Hamilton, & Daily, 2012) but at the same time its perceptual disorderliness may be distressing (Ross, 2000; Tullett et al., 2015), or a certain natural environment may discourage rule-breaking behaviors (Kuo & Sullivan, 2001a, 2001b) but at the same times its perceptual disorderliness may encourage rule-breaking behaviors (Kotabe et al., 2016b; J. Q. Wilson & Kelling, 1982). The net effect may be more consistent with a beneficial "nature response", but there are various possible explanations, only one of which is that the effect of nature trumps the effect of disorder. For example, self-regulatory resources may be restored by nature (S. Kaplan & Berman, 2010), and in turn, may aid in downregulating stress and unwanted impulses (Kotabe & Hofmann, 2015), thus mitigating the behavioral consequences of perceptual disorder in natural environments.

There are also implications for other lines of research. If the high-level scene semantics of nature have strong affective importance tied to them, it may be difficult to build visual-feature-based models that predict cognitive dimensions of these kinds of scenes. For example, models that try to predict memorability of scenes based on global visual features of scenes seem to underestimate memorability of images of higher natural content (Isola, Xiao, Torralba, & Oliva, 2011), perhaps because they do not take into account affect-laden scene semantics. The importance of scene semantics for nature's aesthetics and the generally stronger effects of naturalness (e.g., compared to disorder in our study) could be related to nature's unique ties with dimensions with an evolutionary basis such as survivability (e.g., Nairne, Pandeirada, & Thompson, 2008; Wilson, 1984). This too is an area worthy of further inquiry.

With regard to practical implications, knowledge about people's environmental preferences are weighted into decisions by architects, urban planners, politicians, and other professionals who are responsible for improving the environment. And rightly so—considering that aesthetic preference for natural environments is intimately linked to nature's restorative potential (Han, 2010; Hartig & Staats, 2006; Purcell et al., 2001; Staats et al., 2010; Ulrich, 1983; Van den Berg et al., 2003), perhaps aesthetic preferences should be weighted even more.

The results of this study suggest that both the perception of nature and order are important, as well as paying regard to the low-level visual features that give rise to these percepts. If naturalness and disorder more or less independently affect aesthetic preference, then highly ordered nature scenes (e.g., imagine a Zen garden) should be particularly beautiful. Supporting this prediction, in both Experiments 1a-c and Experiments 2a-c, the most ordered natural scenes were most aesthetically preferred and the most disordered built scenes were least aesthetically preferred, with orderly built scenes and disorderly natural scenes falling between in a nearly linear pattern (see Figure 4). That said, the orderly and natural scenes in these experiments were not manmade like a Zen garden. Zen gardens may be particularly beautiful because of their naturalness and order, but part of their beauty could be attenuated by perceived human influence, perhaps via shifts in perceived naturalness and order. Relevant to this idea is work on "technological nature" (Kahn, 2011), e.g., natural scenes presented via digital screens, which suggests that something important is lost when nature is filtered through such technologies. Generally speaking, the beneficial effects of nature are attenuated by such technologies (Kahn, Severson, & Ruckert, 2009). Therefore, Zen gardens may be very beautiful, but if one were to stumble upon an untouched natural landscape that is highly ordered like a Zen garden, it may be exalted into an aesthetic class of its own. An interesting avenue is to take this idea of aesthetic adulteration via human influence a step further and test other consequences of human influence on aesthetic preference (e.g., changing colors, edges, shapes, etc. of natural entities and environments). Does any human influence adulterate nature's aesthetics, or do certain human influences have negligible or even beneficial effects on nature's aesthetics?

[INSERT FIGURE 4 HERE]

As the world becomes more populated and urbanized, there is a pressing demand to incorporate nature into built environments. Not only does it have aesthetic, psychological, and physical health benefits, it also is economically sensible—according to a report by Booze Allen Hamilton (2015), green construction is predicted to directly contribute $303.4 billion to the U.S. gross domestic product and support 3.9 million jobs in the U.S. from 2015-2018. In addition, as virtual reality (another multibillion-dollar industry) becomes more of a reality, there is a growing interest in designing salubrious virtual environments. This paper suggests that order should be considered in the design of both greenspace environments and virtual environments.

References

Appleton, J. (1996). *The experience of landscape*. Chicago, IL: Wiley.

Arnheim, R. (1974). *Entropy and art: An essay on disorder and order*. London, England: University of California Press.

Bar, M., & Neta, M. (2006). Humans prefer curved visual objects. *Psychological Science*, *17*, 645–648.

Bar, M., & Neta, M. (2007). Visual elements of subjective preference modulate amygdala activation. *Neuropsychologia*, *45*, 2191–2200.

Berens, P. (2009). CircStat: a MATLAB toolbox for circular statistics. *Journal of Statistical Software*, *31*, 1–21.

Berman, M. G., Hout, M. C., Kardan, O., Hunter, M. R., Yourganov, G., Henderson, J. M., … Jonides, J. (2014). The perception of naturalness correlates with low-level visual features of environmental scenes. *PLOS ONE*, *9*, e114572.

Berman, M. G., Jonides, J., & Kaplan, S. (2008). The cognitive benefits of interacting with nature. *Psychological Science*, *19*, 1207–1212.

Berman, M. G., Kross, E., Krpan, K. M., Askren, M. K., Burson, A., Deldin, P. J., … Jonides, J. (2012). Interacting with nature improves cognition and affect for individuals with depression. *Journal of Affective Disorders*, *140*, 300–305.

Bratman, G. N., Hamilton, J. P., & Daily, G. C. (2012). The impacts of nature experience on human cognitive function and mental health. *Annals of the New York Academy of Sciences*, *1249*, 118–136.

Brunswik, E. (1947). Systematic and representative design of psychological experiments. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* (pp. 143–202).

Canny, J. (1986). A computational approach to edge-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8*, 679–698.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*, 1–29.

Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cognitive Brain Research*, *20*, 226–241.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*, 167–192.

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science*, *317*, 1351–1354.

Han, K.-T. (2010). An exploration of relationships among the responses to natural scenes scenic beauty, preference, and restoration. *Environment and Behavior*, *42*, 243–270.

Hartig, T., & Staats, H. (2006). The need for psychological restoration as a determinant of environmental preferences. *Journal of Environmental Psychology*, *26*, 215–226.

Ibarra, F., Kardan, O., Hunter, M. R., Kotabe, H. P., Meyer, F. A., & Berman, M. G. (2017). *How low-level features relate to high-level semantics to predict scene aesthetic preference and naturalness*.

Kahn, P. H. (2011). *Technological nature: Adaptation and the future of human life*. Cambridge, MA: MIT Press.

Kahn, P. H., Severson, R. L., & Ruckert, J. H. (2009). The human relation with nature and technological nature. *Current Directions in Psychological Science*, *18*, 37–42.

Kaplan, R., & Austin, M. E. (2004). Out in the country: Sprawl and the quest for nature nearby. *Landscape and Urban Planning*, *69*, 235–243.

Kaplan, R., & Kaplan, S. (1989). *The experience of nature: A psychological perspective*. Cambridge, MA: Cambridge University Press.

Kaplan, S. (1995). The restorative benefits of nature: Toward an integrative framework. *Journal of Environmental Psychology*, *15*, 169–182.

Kaplan, S., & Berman, M. G. (2010). Directed attention as a common resource for executive functioning and self-regulation. *Perspectives on Psychological Science*, *5*, 43–57.

Kaplan, S., Kaplan, R., & Wendt, J. S. (1972). Rated preference and complexity for natural and urban visual material. *Perception & Psychophysics*, *12*, 354–356.

Kardan, O., Demiralp, E., Hout, M. C., Hunter, M. R., Karimi, H., Hanayik, T., … Berman, M. G. (2015). Is the preference of natural versus man-made scenes driven by bottom-up processing of the visual features of nature? *Frontiers in Psychology*, *6*, 471.

Kosslyn, S. M. (1996). *Image and brain: The resolution of the imagery debate*. Cambridge, MA: MIT press.

Kotabe, H. P., & Hofmann, W. (2015). On integrating the components of self-control. *Perspectives on Psychological Science*, *10*, 618–638.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2016a). Can the high-level semantics of a scene be preserved in the low-level visual features of that scene? *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, *38*, 1721–1726.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2016b). The order of disorder: Deconstructing visual disorder and its effect on rule-breaking. *Journal of Experimental Psychology: General*, *145*, 1713–1727.

Kuo, F. E., & Sullivan, W. C. (2001a). Aggression and violence in the inner city effects of environment via mental fatigue. *Environment and Behavior*, *33*, 543–571.

Kuo, F. E., & Sullivan, W. C. (2001b). Environment and crime in the inner city: Does vegetation reduce crime? *Environment and Behavior*, *33*, 343–367.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621.

Mellet, E., Tzourio, N., Denis, M., & Mazoyer, B. (1995). A positron emission tomography study of visual and mental spatial exploration. *Journal of Cognitive Neuroscience*, *7*, 433–445.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*, 145–175.

Özgüner, H., & Kendle, A. D. (2006). Public attitudes towards naturalistic versus designed landscapes in the city of Sheffield (UK). *Landscape and Urban Planning*, *74*, 139–157.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*, 241–263.

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

Purcell, T., Peron, E., & Berto, R. (2001). Why do preferences differ between scene types? *Environment and Behavior*, *33*, 93–106.

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.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Staats, H., Van Gemerden, E., & Hartig, T. (2010). Preference for restorative situations: Interactive effects of attentional state, activity-in-environment, and social context. *Leisure Sciences*, *32*, 401–417.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391–412.

Ulrich, R. S. (1983). Aesthetic and affective response to natural environment. In *Behavior and the natural environment* (pp. 85–125). New York, NY: Springer.

Van den Berg, A. E., Koole, S. L., & van der Wulp, N. Y. (2003). Environmental preference and restoration: (How) are they related? *Journal of Environmental Psychology*, *23*, 135–146.

Van den Berg, A. E., & van Winsum-Westra, M. (2010). Manicured, romantic, or wild? The relation between need for structure and preferences for garden styles. *Urban Forestry & Urban Greening*, *9*, 179–186.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*, 1115–1125.

Wilson, E. O. (1984). *Biophilia*. Cambridge, MA: Harvard University Press.

Wilson, J. Q., & Kelling, G. L. (1982). Broken windows. *Atlantic Monthly*, *249*, 29–38.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*, 141–145.

Author Contributions: H.P.K., O.K., and M.G.B. designed the experiments. H.P.K. conducted the experiments. O.K. wrote the MATLAB scripts to quantify visual features and construct low-level visual stimuli. H.P.K. conducted the data analysis and interpretation with assistance from O.K. and M.G.B. H.P.K., O.K., and M.G.B. wrote the manuscript.

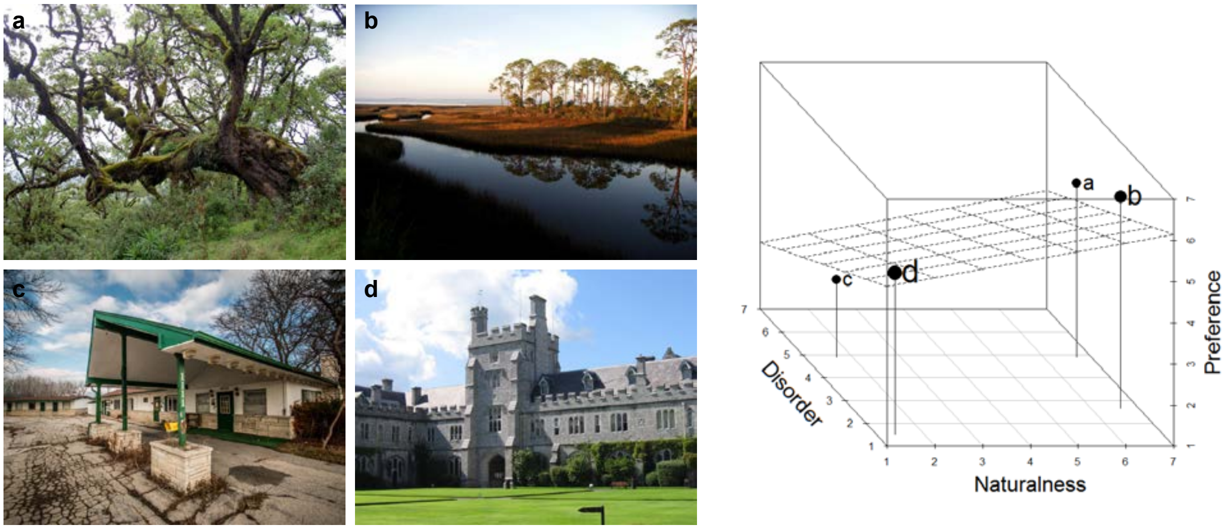

Figure 1. On the left, four scenes from the set of 916 scene images used in Experiments 2a-c that exemplify the coexistence of (a) naturalness and disorder; (b) naturalness and order; (c) builtness and disorder; and (d) builtness and order. On the right, these scenes are mapped in three-dimensional space relative to the regression plane when simultaneously regressing aesthetic preference ratings on naturalness ratings, disorder ratings, and their interaction in this set of experiments.

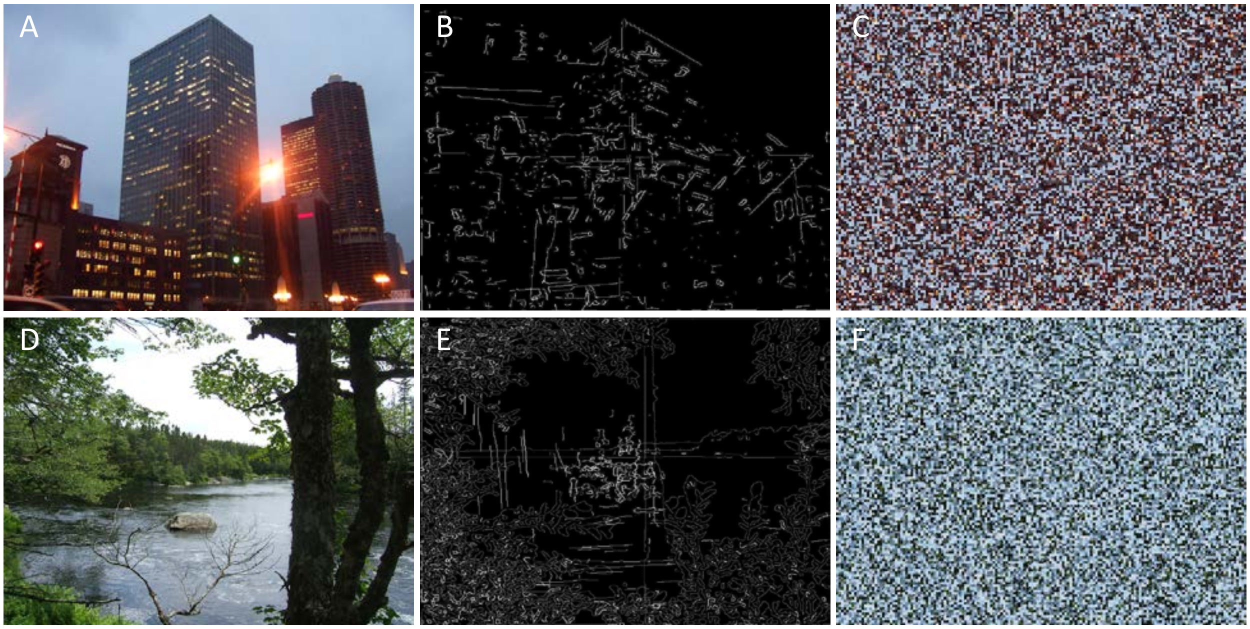

Figure 2. Examples of the highest-rated built and highest-rated natural scene images from the set of 260 scene images and their derived stimuli. (A) Original highly-built scene image (from Experiments 1a-c); (B) its derived scrambled-edge stimulus (Experiments 3a-c); and (C) its scrambled-color stimulus (Experiments 3d-f). (D) Original highly-natural scene image (from Experiments 1a-c); (E) its derived scrambled-edge stimulus (Experiments 3a-c); and (F) its scrambled-color stimulus (Experiments 3d-f).

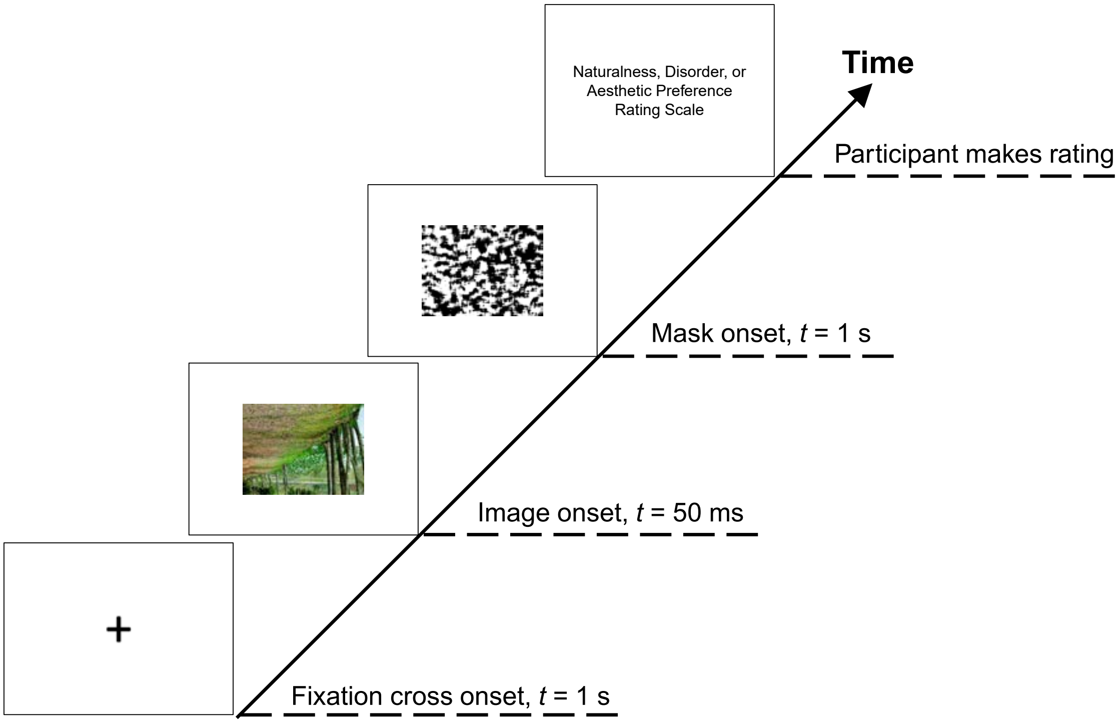
Naturalness, Disorder, or Aesthetic Preference Rating Scale
Time
Participant makes rating
Mask onset, *t* = 1 s
Image onset, *t* = 50 ms
+
Fixation cross onset, *t* = 1 s

Figure 3. A single trial of the rapid-scene-presentation procedure used in Experiments 4a-c. A fixation cross appeared for 1 s. An inverted scene image from the set of 260 scene images was then presented for 50 ms. The scene image was then masked by one of eight perceptual masks. The mask was presented for 1 s. Afterward, participants were prompted to make a rating of naturalness, disorder, or aesthetic preference. Participants were given unlimited time to make a rating. The next trial started automatically after a rating was made.

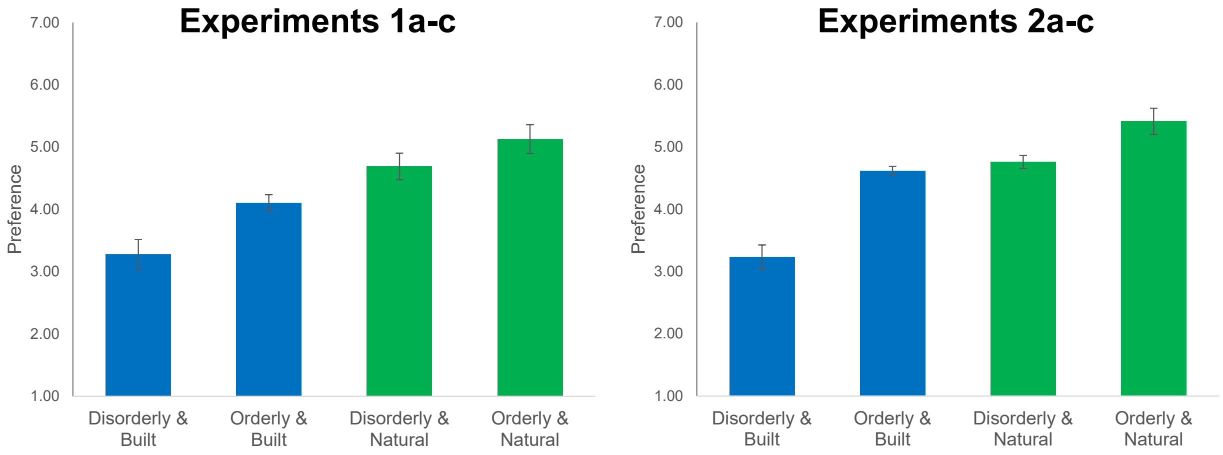

Figure 4. Mean aesthetic preference ratings for scene images rated in the top quintiles of builtness/naturalness and order/disorder in Experiments 1a-c and Experiments 2a-c. Error bars indicate mean±s.e.m.