THE UNIVERSITY OF CHICAGO

DISORDER, NATURALNESS, AND THEIR INFLUENCE ON AESTHETICS AND BEHAVIOR

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE UNIVERSITY OF CHICAGO

BOOTH SCHOOL OF BUSINESS

AND

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

DEPARTMENT OF PSYCHOLOGY

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

HIROKI P. KOTABE

CHICAGO, ILLINOIS

JUNE 2016

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

MY COMMITTEE:

Marc G. Berman (Chair)

Reid Hastie

Wilhelm Hofmann

Daniel Casasanto

SPECIAL THANKS:

My wife, Peggy, and our furry children, Darwin and Javier

My family, especially my dad—maybe one day searching “Kotabe” in Google Scholar will bring up my papers next to yours ;-)

Omid, for the prototyping and things

ABSTRACT

The human has never been and will never be separated from their physical environment. How does the physical environment shape us, our feelings and thoughts, and our behavior? In this dissertation, we present a variety of research motivated by these questions. We focus on two semantic dimensions of the physical environment that are pertinent to this topic—its level of order/disorder and its level of builtness/naturalness. In Chapter 1, we examine whether low-level visual features that influence perceived disorder (*visual disorder cues*) have been overlooked in research on broken windows theory that assumes that social disorder cues and complex social reasoning are necessary for the effect of disorderly environments on rule-breaking behavior. In one set of experiments, we identified key visual disorder cues that generalize across a variety of visual stimuli with a variety of semantic content. In another set of experiments, we demonstrated that visual disorder can encourage cheating. In an additional two experiments, we explored mechanisms of this effect. In Chapter 2, we identify and resolve a paradoxical relationship between disorder and naturalness (if disorder is aesthetically aversive and naturalness is aesthetically pleasing, how is it that natural environments are disorderly?). Across four sets of experiments, we tested three competing hypotheses that could explain this relationship and found that (a) the effects of naturalness and disorder on aesthetic preference are independent and (b) the effect of naturalness on aesthetic preference trumps the effect of disorder. Further, we show that scene semantics are both necessary and sufficient for this nature-trumps-disorder effect, and their interaction with low-level visual features amplifies this effect. In Chapter 3, we examine whether high-level semantics of a scene related to disorder and naturalness can be preserved in the low-level visual features of that scene, contrary to traditional visual perception models that assume

that integration of low-level visual features must occur *before* high-level semantics are perceived. The results of two experiments suggest not only that disorder and naturalness semantics can be preserved in low-level visual features, but also that disorder and naturalness semantics can be preserved in different types of low-level visual features. This research adds to a growing body of literature that suggests that low-level visual features can carry semantic information. Together, these chapters present a variety of evidence substantiating the intimate connection we have to our ever-present physical environments, and provide insight into the linkage between low-level visual processing, semantics, aesthetic sense, and behavior.

GENERAL INTRODUCTION

The human has never been and will never be separated from their physical environment. How does the physical environment shape us, our feelings and thoughts, and our behavior? In this dissertation, we present a variety of research motivated by these questions. We focus on two semantic dimensions of the physical environment that are pertinent to this topic—the degree to which it is perceived as orderly vs. disorderly and the degree to which it is perceived as built vs. natural.

In Chapter 1, we focus on disorderly environments and their behavioral consequences. Disorderly environments are linked to disorderly behaviors. Broken windows theory (Wilson & Kelling, 1982), a highly influential sociological theory of criminal behavior, assumes that social-disorder cues (e.g., litter, graffiti) lead people to reason that they can get away with breaking rules. But what if part of the story is not about reasoning at all? What if *visual disorder* alone is sufficient to encourage rule-breaking? To answer this question, we first conducted a set of experiments (Experiments 1-3) in which we identified key visual disorder cues that generalize across visual stimuli with a variety of semantic content. Our results revealed that spatial features (e.g. non-straight edges, asymmetry) are more important than color features for visual disorder. Exploiting this knowledge, we then reconstructed stimuli with different degrees of visual disorder, absent of social disorder cues, and tested whether visual disorder encourages cheating in a second set of experiments (Experiment 4-6). In these experiments, subtly manipulating visual disorder increased the likelihood of cheating by up to 35% and the average magnitude of cheating by up to 87%. In a final experiment (Experiment 7), we explored some potential mechanisms of these effects using text analysis. This work suggests that explanations for rule-

breaking that assume that *complex social reasoning* is necessary should be reconsidered (e.g., Kelling & Coles, 1997; Sampson & Raudenbush, 2004). Furthermore, these experiments show that simple perceptual properties of the environment can affect complex behavior and sheds light on the extent to which our actions are within our control.

In Chapter 2, we focus on disorder, naturalness, and their joint influence on aesthetic preferences. Exposure to natural environments has various salubrious effects, which have been theoretically and empirically linked to a strong aesthetic preference for such environments. In contrast, exposure to disorderly environments has various detrimental effects, which may be why people find disorderly environments aesthetically aversive. But in our research, we have repeatedly found that natural environments *are* disorderly. What could explain this paradox? We present three competing hypotheses: The effect of naturalness on aesthetic preference trumps the effect of disorder (*nature-trumps-disorder hypothesis*); disorder does not affect aesthetic preference in natural contexts (*harmless-disorder hypothesis*); or disorder has a positive effect on aesthetic preference in natural contexts (*beneficial-disorder hypothesis*). Through a series of experiments, we rule in the nature-trumps-disorder hypothesis and rule out the harmless-disorder and beneficial-disorder hypotheses. In addition, the results of a set of experiments in which we removed scene semantic by extracting and scrambling the low-level visual features of scenes suggests that scene semantics are both *necessary* and *sufficient* for the nature-trumps-disorder effect, and their interaction with low-level visual features amplifies the effect. This suggests that the presence of recognizable entities in an environment can suppress or strengthen the relationships between naturalness, disorder, and aesthetic preference. More generally, this aspect

of this work is relevant to psychological theories concerning the joint influence of lower-level visual features, higher-level semantics, and their interaction on affect and cognition.

In Chapter 3, we focus on whether disorder and naturalness semantics of a scene can be preserved in the low-level visual features of that scene. Scenes of environments contain low-level visual features such as edges and colors and high-level semantic features such as recognizable objects, places, and descriptors. Traditional visual perception models suggest that integration of low-level visual features and segmentation of the scene must occur before high-level semantic features are perceived. This view implies that low-level visual features alone do not carry semantic information. Here we present evidence that suggests otherwise. We show not only that high-level semantics can be preserved in low-level visual features, but also that different high-level semantics can be preserved in different types of low-level visual features. Specifically, the 'disorder' of a scene is preserved in low-level edge features better than low-level color features, whereas the converse is true for 'naturalness.' These findings suggest that semantic processing may start earlier than thought before, and integration of low-level visual features and segmentation of the scene may occur after semantic processing has begun, or in parallel.

CHAPTER 1: THE ORDER OF DISORDER

We feel, think, and act differently when we are in disorderly environments. According to broken windows theory (J. Q. Wilson & Kelling, 1982), a highly influential sociological theory of criminal behavior and rule-breaking (Sampson & Raudenbush, 2004), disorderly environments encourage rule-breaking behaviors that snowball into major communal problems. For example, if we see litter, it encourages us to litter, and through iterations this can lead to a major littering problem. To compound the problem, the effect of disorderly environments on rule-breaking not only spreads within a domain, but also between domains (Keizer, Lindenberg, & Steg, 2008). For example, if we see litter, we may throw a rock through a window. If we see a broken window, we may steal. And so on. Research employing experimental methods and large-sample correlational methods has converged on the idea that disorderly environments can cause disorderly behaviors (Braga et al., 1999; Braga & Bond, 2008; Keizer et al., 2008; Linares et al., 2001). In addition, environmental disorder has been linked to other detrimental outcomes such as perceived powerlessness (Geis & Ross, 1998), distress (Cutrona, Russell, Hessling, Brown, & Murry, 2000), fear of crime and feeling unsafe (Perkins & Taylor, 1996), depression (Ross, 2000), anxiety and performance-monitoring (Tullett, Kay, & Inzlicht, 2015), and self-regulatory failure (Chae & Zhu, 2014; Kathleen D. Vohs, Redden, & Rahinel, 2013) (for a review, see Kotabe, 2014).

The dominant explanations for broken windows phenomena assume complex reasoning about social cues related to rule-breaking (e.g., litter, graffiti, an abandoned building) (*social disorder cues*) (Kelling & Coles, 1997; Sampson & Raudenbush, 2004; Wilson & Kelling, 1982). For example, when seeing such social cues people may reason that policing is low,

misconduct is the norm, or poverty is prevalent (Sampson & Raudenbush, 2004). In their seminal paper, Wilson and Kelling (1982) wrote, "window-breaking does not necessarily occur on a large scale because some areas are inhabited by determined window-breakers whereas others are populated by window-lovers; rather, one unrepaired broken window is a signal that no one cares, and so breaking more windows costs nothing." And in a highly-cited paper on the perception of disorder, Sampson and Raudenbush (2004) asked, "… is disorder filtered through a reasoning based on stigmatized groups and disreputable areas?"

There is a general problem with such perspectives which is that they are based on research that has not taken a nuanced approach to defining and assessing 'disorder', thus any specific interpretations of the evidence are dubious (see Harcourt, 2009). One of the specific issues arising from the general problem is that previous research has confounded visual disorder and social disorder. By 'visual disorder' we mean the perception of disorder that is attributable to low-level visual features (*visual-disorder cues*). For example, Wilson and Kelling (1982) use the term 'disorder' in reference to both physical environments richly varying in low-level visual cues (e.g., imagine litter being present vs. absent) and to social environments that vary less in low-level visual cues (e.g., imagine a drunk person in public vs. the same person sober in public), without making any distinction between the two. Sampson and Raudenbush (2004) attempt to separate a 'physical disorder' component of environmental disorder, but their operationalization (subjective rating to three questions: how much of a problem is litter/trash, graffiti, and vacant housing/storefronts [in your neighborhood]?) also is unclear about the extent of visual-disorder cues vs. social-disorder cues in the environment that are relevant. Even with this ambiguity,

researchers continue to attribute their findings entirely to social-disorder cues while overlooking visual-disorder cues.

Psychological studies bearing on broken windows theory increase control over extraneous variables by using experimental manipulations but still contain the central problem of confounding social-disorder cues and visual-disorder cues. For example, Keizer at al. (2008), utilizing field experimental methods, observed pedestrian littering and stealing behavior in urban environments either containing graffiti or not, without discussing the visual features that systematically varied between conditions. Vohs et al. (2013) and Chae and Zhu (2014), utilized laboratory experimental methods, and manipulated environmental disorder by making their experimental settings messy or tidy. In some of these experiments, the same objects were placed in the rooms in both conditions, thus holding color features more or less constant. However, spatial visual features systematically varied between conditions (e.g., pencils and papers neatly arranged on a table in the tidy condition vs. scattered about on the floor and table in the messy condition), but the potential for visual disorder to be contributing was not discussed. This is not all that surprising, given the absence of theorizing about the role of visual disorder in causing disorderly behaviors.

The natural question, then, is what if all along there has been an unexplored side to these stories that is not about complex reasoning about social-disorder cues? The previous research raises the question of whether visual disorder by itself can encourage rule-breaking, or if broken windows phenomena are truly driven entirely by complex social reasoning. Deconstructing visual disorder and showing that visual-disorder cues, absent of any social-disorder cues,

encourages rule-breaking would shed light on this important yet completely overlooked determinant of rule-breaking behavior (see Figure 1).

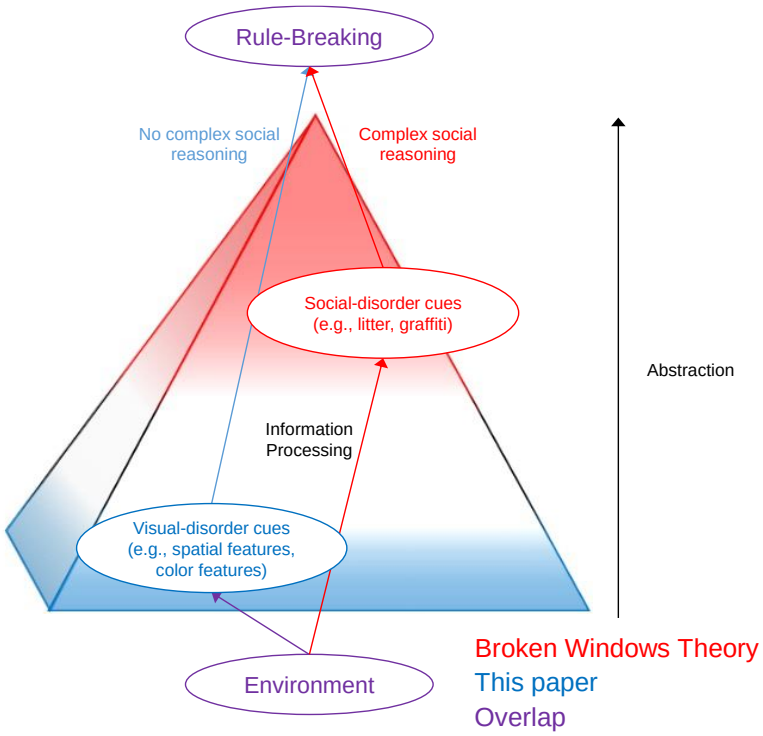


Figure 1: The relation between this chapter and broken windows theory. Broken windows theory proposes that complex reasoning about social-disorder cues (e.g., litter, graffiti, an abandoned building) in the environment encourages rule-breaking behaviors. This chapter is concerned with the possibility that visual disorder cues alone are sufficient to encourage rule-breaking behaviors without invoking such reasoning. Social-disorder cues often have components of visual disorder that we identified in the present research. Consequently, research bearing on broken windows theory often confounds visual-disorder cues and social-disorder cues, and attributes the whole effect to the latter.

By what mechanisms could visual-disorder cues affect complex behaviors such as rule-breaking? To address this, we briefly turn to the vision literature, and expand upon it. The

traditional non-Gestalt perspective (i.e., hierarchical perspective) is that early visual processing of scenes involves the extraction of low-level spatial features (e.g., edges) that are grouped in later stages of visual processing to facilitate recognition (e.g., Biederman, 1987; David Marr, 1976; D. Marr & Hildreth, 1980). Thus, for example, before litter or graffiti is recognized in a scene, the viewer would extract a sort of 'primal sketch' (David Marr, 1976) of the scene which would include spatial features of the litter and graffiti that aid their later recognition. Recent research suggests that such primitive spatial sketches can carry semantic information (Kotabe, Kardan, & Berman, 2016a; Oliva & Torralba, 2006; Walther, Caddigan, Fei-Fei, & Beck, 2009) and interact with higher level goal-driven behavior (Kardan, Henderson, Yourganov, & Berman, in press). In fact, a recent study suggests that decision-making involving interpretation of visual information may occur in visual cortex without involving fronto-parietal regions typically associated with such decision-making (Brascamp, Blake, & Knapen, 2015). These studies do not indicate that visual-disorder cues would necessarily carry semantics associated with disorder (e.g., litter, graffiti, poverty) (though Sampson & Raudenbush, 2004 suggest this could be the case), however it does suggest that semantic processing may play a role. Because semantics are involved in the production of complex behaviors such as self-regulation (e.g., Metcalfe & Mischel, 1999), they could mediate an effect of visual disorder on rule-breaking behavior.

In addition, low-level feature extraction may vary in terms of processing difficulty (Field, 1987; Kinchla, 1977; Olshausen & Field, 1996; Witkin & Tenenbaum, 1983) independent of semantics. Witkin and Tenenbaum argue that a key process in visual perception is the organization of visual information into coherent and manageable chunks (see also Mahoney, 1987). Kinchla's research suggests that structural redundancy between higher and lower forms

(e.g., as found in symmetry) aids in efficiently making sense of visual information (see also Field, 1987; Olshausen & Field, 1996). These studies suggest that processing visually disordered stimuli may be more difficult than processing visually ordered stimuli. Because processing difficulty is involved in the production of complex behaviors such as high-level decision-making (for a review, see Alter & Oppenheimer, 2009), it may also mediate an effect of visual disorder on rule-breaking behavior.

Taken together, this research supports that even low-level visual features, such as those that define visual disorder and visual order, have the potential to affect complex behaviors such as rule-breaking. Here, we focus on this broader question to determine whether there is truth to our concern that visual-disorder cues and social-disorder cues are confounded in previous research bearing on broken windows theory, and whether the former can have an effect in isolation. To answer these questions, our first goal was to define visual disorder objectively by identifying specific visual-disorder cues. Our second goal was to examine if exposure to these cues can encourage rule-breaking behavior. Towards the first goal, we took a principled approach to quantifying, extracting, and scrambling objective visual features of various environmental scenes (as advocated by Geisler, 2008) and analyzed hundreds of people's disorder ratings for various relevant stimuli. Towards the second goal, we conducted behavioral experiments using large and diverse online samples to investigate whether visual-disorder cues alone could encourage cheating behavior.

Overview of Experiments

Across all of our experiments, we sampled broadly from real-world environments by utilizing 260 images of environmental scenes that ranged from more urban to more natural

according to ratings previously collected in the laboratory (Berman et al., 2014; Kardan et al., 2015) (see Figure S1 for examples; all images can be downloaded here in original resolution: goo.gl/S8ShgT).[1] Experiments 1-3 focused on identifying visual-disorder cues. Together, these are the first experiments to define subjective visual disorder at an objective and quantifiable level. They address questions such as: What is visual disorder and how can it be manipulated? What quantifiable visual features of a scene can be used to estimate its level of subjective disorder? Such methodological and empirical knowledge would be useful for theory-building, as well as design. In Experiment 1, we used the images in their original form. In Experiments 2 and 3, we extracted and scrambled low-level visual features of the images to remove the possibility of confounding social-disorder cues. Experiments 4-6 focused on testing the broader hypothesis pertinent to broken windows theory that visual-disorder cues alone, absent of social-disorder cues, are sufficient to encourage rule-breaking. In these experiments, we reconstructed visual disorder based on what was learned from deconstructing visual disorder in Experiments 1-3. Lastly, in Experiment 7, we had participants write about the visually ordered and visually disordered stimuli to inquire about what these stimuli caused people to overtly think about. This would further rule out the possibility that social-disorder cues were still present in Experiments 4-6. Experiment 7 also provided some preliminary information regarding the mechanisms by which visual disorder causes cheating behavior.

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

Experiment 1: Quantifying Visual Disorder Part I

We had people rate the scene images in terms of disorder. In addition, we quantified spatial- and color-related low-level visual features of these images (as in Berman et al., 2014; Kardan et al., 2015) to test the extent to which they predicted perceived disorder. Spatial features included non-straight edge density, straight edge density, and asymmetry. Color features included hue, saturation, value, and the standard deviations of those measures as measures of hue diversity, saturation diversity, and value diversity. As aforementioned, the results of Vohs et al. (2013) and Chae and Zhu (2014) suggest that color features may be less important for visual disorder than spatial features. This experiment is the first test of this possibility. Confirming this hypothesis would have the additional benefit of considerably reducing the dimensions of the visual feature space necessary to construct visually disordered stimuli.

Method

Participants and design. 105 US-based adults (51 men; 54 women) were recruited from the online labor market Amazon Mechanical Turk (AMT). Sample size and stopping rule were based on our goal to receive ~20 ratings per image. Ages ranged from 19 to 75 (M = 36.15, SD = 12.07). 84 participants identified primarily as White/Caucasian, 10 as Black/African American, 5 as Hispanic/Latino, 4 as Asian/Asian American, and 2 as Native Hawaiian/Pacific Islander. The median experiment duration was 5 minutes and 58 seconds and participants were compensated $0.50 for participating. Informed consent was administered by the Institutional Review Board (IRB) of the University of Chicago.

Procedure. Participants agreed to a consent form that only revealed high-level information about the purpose of the study. They were then taken to an instruction page where

they were told that they would be presented a series of 50 images of various environmental scenes and that they were to rate each scene in terms of how disorderly or orderly it looked. Here and in Experiments 2 and 3, we did not explicitly define disorder because our goal was to evaluate systematic relationships between low-level visual features and people's subjective disorder ratings.

Next, they were taken to the image rating task (IRT). Scene images (all 4:3 ratio) were presented on a plain white background in a 600 x 450 pixel frame. Just below the image frame, text was presented that asked, "How disorderly or orderly is this environment?" And just below that was a seven-point semantic differential scale anchored by the options "very disorderly" and "very orderly." Each participant was randomly presented 50 of the 260 scene images. The randomization scheme had two layers. First, we randomly selected 10 images from each quintile of urbanness/naturalness to ensure diverse semantic content. Second, we presented these 50 images in random order, thus each participant viewed images containing a wide sample of scene types. Immediately after making a rating, they would automatically proceed to the next image until all 50 images were rated. Here and in Experiments 2 and 3, presentation time was not fixed in order to assess spontaneous disorder ratings.

Quantifying spatial and color features. We utilized MATLAB's Image Processing Toolbox to quantify three low-level spatial features and six low-level color features to statistically estimate how much perceived disorder was due to objective spatial and color features of the scene. The spatial features we 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), and vertical reflectional asymmetry ("asymmetry"

for short; a measure of how well the left and right halves of the image mirror each other). The resulting color features, based on the standard Hue-Saturation-Value (HSV) model, were: hue (a measure of the average color appearance of a scene), saturation (a measure of how pure and intense the colors of the scene are on average), and value (a measure of the average luminance of a scene). We also used standard deviations of those color measures as measures of hue diversity, saturation diversity, and value diversity. Straight edge density and non-straight edge density and saturation, mean value, SD of saturation, and SD of value were all quantified from their respective maps created as in (Berman et al., 2014; Kardan, Demiralp, et al., 2015). Because hue of a pixel is an angular value, mean and SD of hue of the images 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 of the edge map of images. These sums were then normalized to [0 1] range by being divided by the total number of non-zero pixels in the edge map of the corresponding image (total edge space). See Table S1 (in appendix) for a correlation matrix of all these visual features.

Results

Data analysis was conducted on image-level summary statistics. We regressed mean disorder ratings on all of the individual spatial and color features. About a fifth of the variance in mean disorder ratings was explained by these visual features, $R^2_{adj} = .17$. Non-straight edge density had the largest effect (see Table 1). A linear contrast indicated that the average effect of the spatial features was significantly larger than the average effect of the color features, $F = 11.46$, $p = .001$. To compare variance in disorder ratings explained by spatial vs. color features, we also separately regressed mean disorder ratings on the spatial features and the color features

(see supplementary materials for results). Adjusting for the number of predictors, the spatial features explained over ten times as much variance as the color features, $R^2_{adj} = .10$ vs. $R^2_{adj} < .01$.

Table 1: Low-Level Visual Features Predicting Disorder Ratings in Experiment 1

Predictor	β	SE	P	η_p^2
Spatial				
Non-straight edge density	0.74	0.11	< .001	.150
Straight-edge density	0.17	0.08	.006	.019
Asymmetry	0.21	0.09	.004	.021
Color				
Hue	-0.12	0.07	.024	.013
Saturation	-0.20	0.08	.002	.024
Value	0.04	0.06	.587	.002
SD hue	0.16	0.08	.089	.019
SD saturation	0.06	0.08	.458	.002
SD value	0.07	0.06	.456	.004

These results begin to corroborate that spatial features—particularly non-straight edge density—are more important for visual disorder than are color features. However, because the scene images not only contained visual cues but also possible social cues, we did not have full control over whether the latter influenced the results. In the following experiments, we extracted and scrambled low-level visual features from the scene images to remove social cues. Thus, in the following experiments the possibility of confounding visual-disorder cues and social-disorder cues was substantially reduced.

Experiment 2: Quantifying Visual Disorder Part II

We separately extracted and scrambled the edge features and the color features from the scene images to remove social cues while preserving low-level visual features of the scenes (see Figure 2b-c). People were randomly assigned to rate these new scrambled-edge or scrambled-

color stimuli in terms of disorder as in Experiment 1. With these ratings, we could statistically estimate the extent to which perceived disorder at the scene level was a function of visually-disordered edges vs. visually-disordered colors. Based on the results of Experiment 1, we predicted that scene-level disorder would be better predicted by visually-disordered edges than visually-disordered colors.

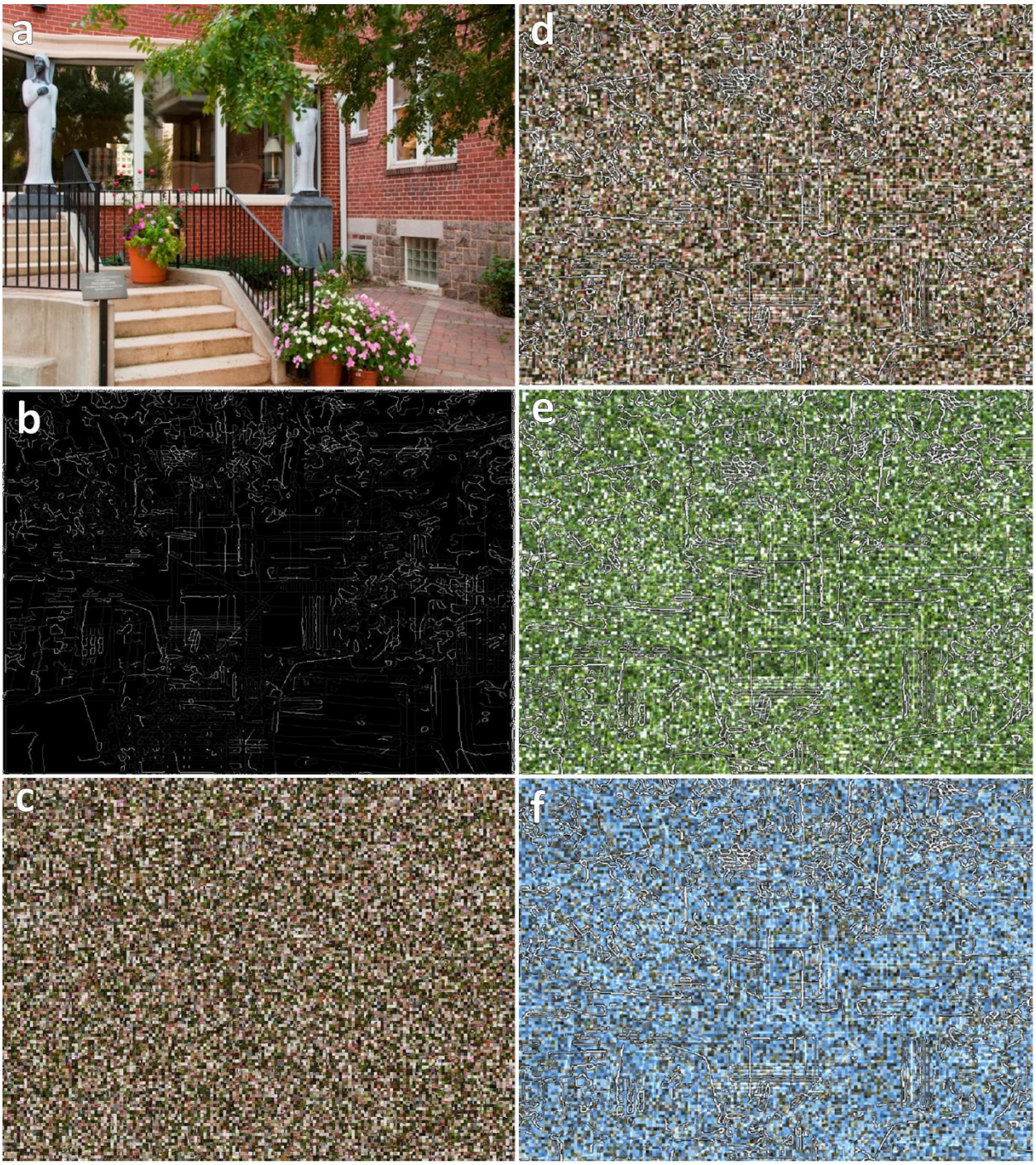


Figure 2: One sample scene image and its derived stimuli. (a) The original image used in Experiment 1; (b) its scrambled-edge stimulus and (c) scrambled-color stimulus used in Experiment 2; and (d) its color-congruent stimulus, (e) color-incongruent stimulus, and (f) control stimulus used in Experiment 3. All images can be downloaded here in original resolution: goo.gl/S8ShgT

Method

Participants and design. 191 US-based adults (108 men, 82 women, 1 other) were recruited from AMT and participated in this two-condition (stimuli: scrambled-edge stimuli vs. scrambled-color stimuli) between-subjects experiment. Sample size and stopping rule were based on our goal to receive ~20 ratings per image. Ages ranged from 18 to 64 (M = 32.16, SD = 11.09). 159 participants identified primarily as White/Caucasian, 11 as Asian/Asian American, 10 as Black/African American, 8 as Hispanic/Latino, and 3 as other. The median experiment duration was 4 minutes and 16 seconds and participants were compensated $0.50 for participating. Informed consent was administered by the IRB of the University of Chicago.

Creating scrambled-edges and scrambled-color stimuli. For the scrambled-edge stimuli, we devised a method to remove possible social cues while preserving edge formations from the original scene images as much as possible (see Figure S2 for an illustration of the processes of this method). First, a mask matrix was 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 (process 1 in Figure S2). 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 half of every mask having, on average, half a surface of 1s and half a surface of 0s. Next, the edge map of the target image (process 2 in Figure S2), created as in (Berman et al., 2014; Kardan et al., 2015), was randomly rotated either 90 or 270 degrees and overlaid on the 180-degrees-rotated edge map (process 3 in Figure S2), creating a stimulus comprising twice as many edges (but same straight and non-straight edge

ratios) as the scene image. This stimulus was then multiplied (dot product) by the mask so that half of its edges got removed at random (process 4 in Figure S2). 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.

For the scrambled-color stimuli, we randomly repositioned windows of 5*5 pixels from the image. The window size was selected so that (a) social cues would become non-discernable, and (b) the color textures of the scene would be preserved. For example, using 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 parts of the scene identifiable, thus possibly preserving social cues.

Procedure. The procedure was the same as in Experiment 1 except that participants in Experiment 2 were randomly assigned to rate a random 50 of the 260 scrambled-edge stimuli or a random 50 of the 260 scrambled-color stimuli.

Results

Data analysis was conducted on image-level summary statistics as in Experiment 1. We separately regressed disorder ratings for the scene images (collected in Experiment 1) on disorder ratings for the scrambled-edge and scrambled-color stimuli (collected in Experiment 2) (see Table S2 for a correlation matrix of disorder ratings for these three sets of stimuli). Disorder ratings for the scrambled-edge stimuli significantly predicted disorder ratings for the scene images, $\beta = 0.38$, $p < .001$ ($R_{adj}^2 = .144$). In contrast, disorder ratings for the scrambled-color stimuli did *not* significantly predict disorder ratings for the scene images, $\beta = 0.02$, $p = .731$ ($R_{adj}^2 = .00$). The adjusted R^2s in this study were similar to the adjusted R^2s when separately

regressing disorder ratings on the spatial features alone (R^2_{adj} = .124) and the color features alone (R^2_{adj} = .038) in Experiment 1, supporting our method of quantifying visual features in Experiment 1.

These results further corroborate that visual disorder is more a function of spatial features than color features. However, the relative predictive validities of edges vs. colors for perceived disorder was tested statistically rather than experimentally. To test this experimentally, we created stimuli that pitted edges and colors against each other in Experiment 3.

Experiment 3: Quantifying Visual Disorder Part III

We manipulated the color features while holding the edge features constant (see Figure 2d-f): In a "color-congruent" condition, scrambled edges from disorderly (orderly) scenes were paired with scrambled colors from disorderly (orderly) scenes. In a "color-incongruent" condition, scrambled edges from disorderly (orderly) scenes were paired with scrambled colors from orderly (disorderly) scenes. In a control condition, scrambled edges were paired with scrambled colors from randomly-selected scenes. People were assigned to rate images from all three conditions in terms of disorder. This data allowed us to experimentally test the competition between disorderly edges and disorderly colors in determining visual disorder. If edge features are more important than color features in determining visual disorder the disorder ratings for the color-congruent and color-incongruent stimuli should correlate to a similar degree with the disorder ratings for the original scene images (i.e., color congruency would not affect disorder ratings).

Method

Participants and design. 222 US-based adults (111 men, 111 women) were recruited from AMT and participated in this three-condition (stimuli: control vs. color-congruent stimuli vs. color-incongruent stimuli) within-subjects experiment. Sample size and stopping rule were based on our goal to receive ~20 ratings per image. Ages ranged from 19 to 76 (M = 35.41, SD = 11.31). 178 participants identified primarily as White/Caucasian, 17 as Asian/Asian American, 13 as Black/African American, 11 as Hispanic/Latino, 2 as multiple ethnicities, and 1 as Native American. The median experiment duration was 5 minutes and 58 seconds and participants were compensated $0.75 for participating. Informed consent was administered by the IRB of the University of Chicago.

Creating color-congruent and color-incongruent (and control) stimuli. To create the "scrambled-edges on scrambled-colors" stimuli we overlaid the previously made scrambled-edge stimuli on the scrambled-color stimuli. First, the scrambled-edge images were sorted in order of visual disorder, i.e., image 1, 2, … , k, … , 260, with image 1 being the most visually ordered and image 260 being the most visually disordered. Then the scrambled-edge stimulus made from k_{th} image was overlaid on the scrambled-color stimulus that was (a) made from the same image (color-congruent stimuli), (b) made from the $(260 - k)_{th}$ image (color-incongruent stimuli), or (3) made from j_{th} image where j is a random number between 1 and 260 without resampling (control stimuli).

Because in the resulting stimuli some of the scrambled edges were not discernable from the background scrambled colors, we made one pixel surrounding the edges (which are white) black to preserve the contrast and consistency of the edges. We note that although this did

inevitably remove some color information, the proportion of remaining pixels belonging to color features (~82% on average) was still much larger than the proportion of remaining pixels belonging to edge features (~18% on average) (see Figure 2d-f), making our test of the "edges > colors" hypothesis more conservative.

Procedure. The procedure was the same as in Experiment 2 except for the following differences: Instead of a between-subjects design, participants in Experiment 3 were assigned to all three visual conditions within-subjects. They rated a total of 75 images—25 of the 260 new color-congruent images, 25 of the 260 new color-incongruent images, and 25 of the 260 new control images. The randomization scheme was similar to that used in Experiments 1 and 2: First, we randomly selected five images from each disorder quintile (based on disorder ratings for the scene images), and repeated this for each of the three sets of color stimuli, resulting in 25 images from each set. Second, we presented these 75 images in random order.

Results

Data analysis was conducted on image-level summary statistics as in the previous experiments. To determine whether edge features are more important than color features for visual disorder, we compared the disorder ratings for the three new sets of experimental stimuli to the disorder ratings for the scene images (collected in Experiment 1) (see Table S3 for a correlation matrix of disorder ratings for these four sets of stimuli). If edges were more important than color for visual disorder at the scene level, we would expect similar correlations between disorder ratings for each set of color-manipulated stimuli and disorder ratings for the original scene images. This was indeed the case (see Figure 3). Williams' t-tests (1959a) confirmed that none of the pairwise dependent correlations significantly differed. In addition, a repeated-

measures GLM indicated that stimuli set (color-congruent vs. color-incongruent vs. control) did not significantly interact with scene-disorder quintile on disorder ratings for the color-manipulated stimuli.

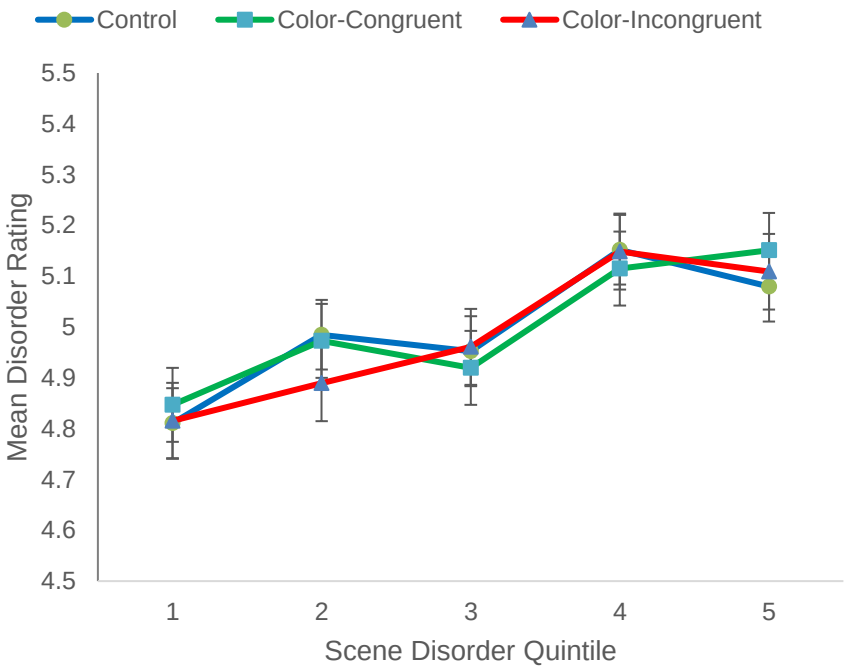


Figure 3: Mean disorder ratings by scene disorder quintile in Experiment 3. The x-axis indicates the quintiles of scene disorder ratings (collected in Experiment 1) on which the color manipulation in Experiment 3 was based. The overlapping lines show that manipulating color features had little to no effect on disorder ratings compared when competing with edge features. Error bars indicate mean ± s.e.m.

These results provide strong evidence that, when competing, edge features are more important for visual disorder than are color features. The three experiments so far each shed light on some quantitative visual features that define visual disorder. They converge on the idea that spatial features such as non-straight edge density are more important than color features for visual disorder. These results not only help us start to define visual disorder, but they also gave

us something tangible to work with to manipulate visual disorder, which was pivotal for the following experiments. In the following experiments, we switch focus to conduct the first investigation into whether being exposed to visual disorder cues alone (as defined in Experiments 1-3) is sufficient to encourage a rule-breaking behavior (i.e., cheating), despite the absence of social disorder cues and complex social reasoning. These experiments bear on the original question we had of whether the confounding of visual disorder cues and social disorder cues in previous research bearing on broken windows theory is problematic.

Experiment 4: The Effect of Visual Disorder on Rule-Breaking Part I

For the following three experiments, we adapted a procedure that (Mazar, Amir, & Ariely, 2008) developed to study one major form of rule-breaking—cheating. Because cheating involves a motivational conflict between what one is tempted to do (i.e., cheat) and what one should do (i.e., not cheat) according to a normative rule, we thought that cheating is a topic particularly suited for a study of rule-breaking more generally. The experimental procedure involves taking a challenging incentivized test, and then later grading oneself, which provides an opportunity to cheat to varying degrees. Immediately before grading themselves, participants were randomly assigned to view and rate visually disordered stimuli or visually ordered stimuli—created based on what we learned from Experiments 1-3—for five minutes (see Figure 4 for examples). In the self-grading phase, participants were told they would receive bonus money for each question they reported as correct. We predicted that people exposed to the visually disordered stimuli would cheat more than people exposed to visually ordered stimuli. In Experiment 4, the cheating incentives were relatively low. In Experiments 5 and 6, we tested the effects of increasing the cheating incentives under high and low disorder salience.

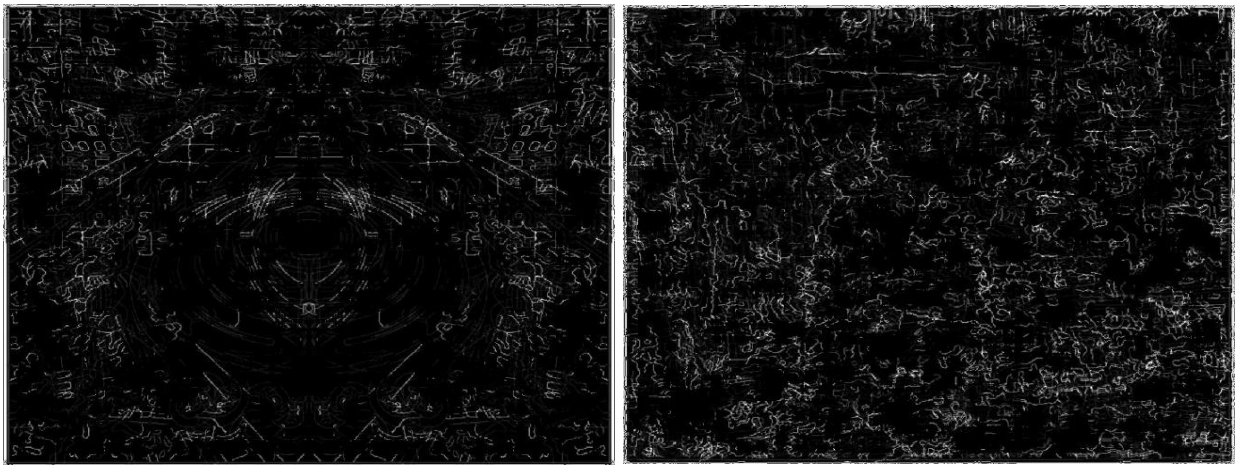


Figure 4: Examples of visual order (left) and visual disorder (right) stimuli. These were constructed based on the results of Experiments 1-3, and used in Experiment 4-6.

Method

Participants and design. 404 US-based adults (180 men, 168 women, 3 other[2]) were recruited from AMT and participated in this two-condition (visually disordered stimuli vs. visually ordered stimuli) between-subjects experiments. Sample size and stopping rule were based on our goal to obtain a large sample size of ~200 per condition to increase power to detect an effect of unknown size. Ages ranged from 18 to 73 (M = 35.94, SD = 11.89). 284 participants identified primarily as White/Caucasian, 20 as Asian/Asian American, 20 as Hispanic/Latino, 19 as Black/African American, 7 as multiple ethnicities, and 1 as Native American. The median experiment duration was 14 minutes and 16 seconds and participants were compensated $0.50 for participating plus a bonus of up to $1.00 (i.e., $0.10 for each correct math answer). Informed consent was administered by the IRB of the University of Chicago.

[2] Demographics for Experiment 4 were collected in a second session that 61 participants did not return to.

Creating visual disorder and visual order. We exploited what we learned in the previous experiments by creating a 2 (symmetry vs. asymmetry) × 2 (visually-ordered edges vs. visually-disordered edges) set of stimuli. This was done by the same method we described for Experiment 2 (Figure S3). This method results in a random mask having, on average, half a surface of 1s and half a surface of 0s. The only difference here is that to artificially make symmetry or asymmetry, we created two random masks for each of the images instead of one. Next, the edge map of the target image was either multiplied (dot product) by both of the masks and then the two resulting matrices were overlaid (asymmetrical stimuli), or was multiplied by only one of the masks and then the resulting matrix was overlaid on the flipped version of itself (symmetrical stimuli). The resulting stimuli had, on average, the same amount of edges with similar edge types, and were either symmetrical or asymmetrical.

Pretesting visual disorder and visual order. To pretest our stimuli, we conducted a 2 (symmetry vs. asymmetry) × 2 (visually-ordered edges vs. visually-disordered edges) within-subjects experiment on AMT with 222 participants. Participants were randomly presented 40 images total (10 from each cell) (see Figure S3 for examples). They rated each image in terms of disorder the same way as in the previous experiments.

Asymmetry and disorderly edges were dummy-coded. A multiple linear regression (R^2_{adj} = .854) indicated that asymmetry, $\beta = 0.78$, $SE = .027$, $p < .001$, $\eta_p^2 = .625$, and disorderly edges $\beta = 0.49$, $SE = .027$, $p < .001$, $\eta_p^2 = .810$, independently increased disorder ratings (see Figure S4). There was no significant interaction.

Manipulating visual disorder and visual order. The 30 most (dis)orderly stimuli were used for our manipulation of visual disorder vs. visual order. We also used the next 5 most (dis)orderly stimuli for the filler task.

Procedure. Participants first received a consent document which disguised the purpose of the experiment by describing it as about the "interplay between visual perception and cognitive performance." Participants then were given a brief introduction to the IRT as in Experiments 1-3. Next, they performed a filler task in which they were presented the 10 filler stimuli for 10 seconds each (one minute and forty seconds of total exposure) and were asked to rate disorder the same way as in the previous experiments. The filler task had two purposes: (1) getting participants acquainted with the IRT before implementing the manipulation and (2) masking the purpose of the study by displaying images both before testing and before self-grading. In the next part of the study, we adapted the procedure that (Mazar et al., 2008) developed to study cheating behavior. This procedure involves taking an incentivized test, then grading oneself on this test, which gives participants the opportunity to cheat. First, in the test phase, participants attempted a task in which they were given two minutes to search for pairs of numbers that add to 10 within 4 × 3 matrices composed of numbers between 0 and 10 with two decimal digits ("Matrices Test"). There were 10 matrices with each containing one solution. Participants were told that they would receive a $0.10 bonus (~4 minutes of work at the median reservation wage on AMT, Horton, Rand, & Zeckhauser, 2011) for each matrix they solved correctly. After two minutes, they were automatically taken to the next part of the study in which we implemented our manipulation. Participants were randomly assigned to view and rate either the 30 visually disordered stimuli or the 30 visually ordered stimuli in terms of disorder. Each

image was presented for 10 seconds (five minutes total exposure). Next, they moved on to the self-grading phase of the experiment. Participants were then instructed to grade themselves on the Matrices Test they had performed earlier. They were reminded that they would be paid $0.10 for each question that they had solved correctly, and that we would take their word for it (i.e., participants could report getting more correct than they actually did). Each matrix from earlier was presented with the correct solution clearly indicated and their answers from before were presented just below each matrix. For each matrix, they were asked to simply respond "Yes" or "No" to the question, "did you get it right?" (see screenshot in supplemental materials). After grading themselves, all participants completed the state PANAS scale (Watson, Clark, & Tellegen, 1988) and a demographics survey before being debriefed. Statistically adjusting for state positive and negative affect did not change the pattern of results in Experiments 4-6, so it is not discussed further.

Results

Manipulation check. An independent-samples t-test confirmed that the manipulation had a significant effect on disorder ratings, $t(402) = 7.88$, $p < .001$, $d = 0.78$, with the visually disordered stimuli ($M = 5.24$, $SD = 0.91$) receiving higher disorder ratings than the visually ordered stimuli ($M = 4.59$, $SD = 0.77$). We note that the mean difference of 0.65 units on a 1-7 scale suggests our manipulation of visual disorder was fairly subtle, which is consistent with our finding from Experiment 1 that visual disorder cues explained about a fifth of the variance in disorder ratings.

Cheating analysis: actual performance vs. reported performance. Cheating was assessed with three a priori tests and one post hoc test. Three participants (< 1% of the sample)

were excluded from the cheating analysis for performing perfectly on the Matrices Test since it would be impossible for them to cheat. First we examined actual performance vs. reported performance in the visual-order vs. visual-disorder condition. Actual performance and reported performance were imperfectly correlated at $r = .44$ indicating that the procedure encouraged people to cheat. We utilized the *lme4* package in R to conduct a linear mixed-model with performance on the Matrices Test predicted by visual condition, actual vs. reported, and their interaction as fixed factors and a random intercept for each participant. Degrees of freedom was estimated with Satterthwaite's approximation. This model revealed a significant main effect of actual vs. reported, $t = 11.48$, $p < .001$, with participants across visual conditions reporting 54% higher performance ($M = 4.71$, $SD = 3.11$) than their actual performance ($M = 3.05$, $SD = 2.06$) on the Matrices Test. However, the interaction between actual vs. reported and visual condition was not significant, $t = 0.44$, $p = .662$, and neither was the simple effect of visual condition within reported performance, $t = 0.29$, $p = .769$.

Cheating analysis: likelihood of cheating. Second, we tested whether the likelihood of cheating differed between the visual order and visual disorder conditions. A chi-square test of independence conducted on a condition-by-cheating (yes/no) contingency table was marginally significant, $\chi^2(1, N = 401) = 3.01$, $p = .083$, $\phi = 0.087$, OR = 1.43, with 44% of participants cheating in the visual-disorder condition and 36% of participants cheating in the visual-order condition (24% relative increase, adjusted residual = 1.73).

Cheating analysis: magnitude of cheating. Third, we compared the visual-order group and the visual disorder group on a measure of absolute cheating magnitude (reported performance – actual performance). There was a descriptive but nonsignificant difference in the

predicted direction, $t(399) = 0.44$, $p = .663$, $d = 0.04$, with those in the visual-disorder condition ($M = 1.72$, $SD = 2.89$) cheating by an 8% larger magnitude than those in the visual-order condition ($M = 1.59$, $SD = 2.89$). The fact that there were 24% more cheaters in the visual-disorder condition, but only 8% increase in magnitude in cheating, means that these extra cheaters tended to cheat by only a little bit.

Cheating analysis: levels of cheating. We conducted a post hoc test of this possible interaction with a multinomial logistic regression model with visual condition predicting each possible level of cheating (cheating by 1 … 10). Non-cheaters were used as the reference category because they were most frequent. There was a significant effect of visual condition on minor cheating (dishonestly adding one point to one's score), $B = 0.65$, $SE = 0.29$, Wald $\chi^2 = 5.05$, $p = .025$, with there being more minor cheaters in the visual-disorder condition (20% minor cheaters) than the visual-order condition (12% minor cheaters) (68% relative increase). There were no significant effects on other levels of cheating.

Experiment 5: The Effect of Visual Disorder on Rule-Breaking Part II

Experiment 5 was identical to Experiment 4 except that we doubled the cheating incentives. We assumed that increasing cheating incentives would increase the temptation to cheat. Thus, we predicted that the effect of visual disorder on cheating would be stronger than in Experiment 4.

Method

Participants and design. 405 US-based adults (206 women, 198 men, 1 unreported) were recruited from AMT and participated in this two-condition (visually disordered stimuli vs. visually ordered stimuli) between-subjects experiments. Sample size and stopping rule were the

same as in Experiment 5. Ages ranged from 19 to 69 (M = 35.34, SD = 11.05). 314 participants identified primarily as White/Caucasian, 29 as Hispanic/Latino, 25 as Black/African American, 22 as Asian/Asian American, 12 as multiple ethnicities, and 3 as Native American. The median experiment duration was 14 minutes and 33 seconds and participants were compensated $0.50 for participating plus a bonus of up to $2.00 (i.e., $0.20 per correct answer), as opposed to $1.00 in the previous experiment. Informed consent was administered by the IRB of the University of Chicago.

Results

Manipulation check. An independent-samples t-test confirmed that the manipulation had a significant effect on disorder ratings, $t(403) = 8.17$, $p < .001$, $d = 0.82$, with the visually disordered stimuli (M = 5.20, SD = 0.94) receiving higher disorder ratings than the visually ordered stimuli (M = 4.53, SD = 0.67). As in Experiment 4, the manipulation of visual disorder was fairly subtle (0.67 units on a 1-7 scale).

Cheating analysis: actual performance vs. reported performance. Cheating was assessed with the three a priori tests conducted in Experiment 4. Six participants (1.5% of the sample) were excluded from the cheating analysis for performing perfectly on the Matrices Test since it would be impossible for them to cheat. Actual performance and reported performance were imperfectly correlated at r = .54 indicating that the procedure encouraged people to cheat. A linear mixed-model with performance on the Matrices Test predicted by visual condition, actual vs. reported, and their interaction as fixed factors and a random intercept for each participant revealed a significant main effect of actual vs. reported, t = 11.10, $p < .001$, with participants across visual conditions reporting 55% higher performance (M = 4.60, SD = 3.27)

than their actual performance (M = 2.97, SD = 2.14) on the Matrices Test. Importantly, there was a significant interaction between actual vs. reported and visual condition, t = 3.59, p < .001, with participants in the visual-disorder condition reporting 70% higher performance than their actual performance and participants in the visual-order condition reporting 39% higher performance than their actual performance (see Figure 5a). The simple effect of visual condition within reported performance was also significant, t = 4.13, p < .001. A follow-up test of multivariate simple effects of actual vs. reported performance within the visual-disorder and visual-order conditions revealed that the effect size in the visual-disorder condition, η_p^2 = .232, was nearly three times larger than the effect size in the visual-order condition, η_p^2 = .078. These results suggest that those in the visual-disorder condition cheated more than those in the visual-order condition.

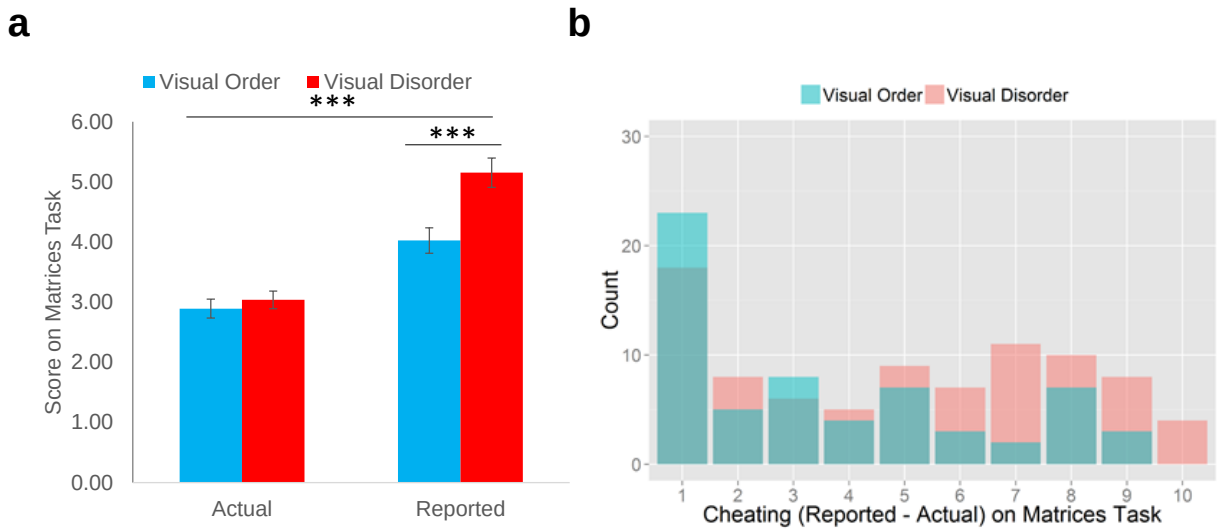


Figure 5. Visual disorder encourages cheating (Experiment 5). (a) Actual vs. reported performance by condition in Experiment 5. Error bars indicate mean ± s.e.m. (b) Magnitude of cheating (reported performance – actual performance) in Experiments 5. *** p < .001

Cheating analysis: likelihood of cheating. A chi-squared test of independence conducted on a condition-by-cheating (yes/no) contingency table was significant, $\chi^2(1, N = 399) = 5.27$, $p = .022$, $\phi = 0.115$, OR = 1.62, with 43% of participants cheating in the visual-disorder condition and 32% of participants cheating in the visual-order condition (35% relative increase, adjusted residual = 2.30). To compare this result to that observed in Experiment 4, we took the difference of the natural logarithm of the ORs (odds ratios), δ, and calculated *SE* of δ with $\sqrt{SE(\ln(\mathrm{OR}_1))^2 + SE(\ln(\mathrm{OR}_2))^2}$. We then obtained *z* with δ / *SE*(δ). The result of the chi-squared test in Experiment 5 did not significantly differ from the chi-squared test conducted in Experiment 4, $\delta = 0.12$, $z = 0.42$, $p = .672$.

Cheating analysis: magnitude of cheating. An independent samples t-test revealed a significant effect of visual disorder on magnitude of cheating, $t(397) = 3.60$, $p < .001$, $d = 0.36$, with those in the visual-disorder group ($M = 2.12$, $SD = 2.12$) cheating by 87% larger magnitude relative to those in the visual-order condition ($M = 1.13$, $SD = 2.26$) (see Figure 5b). To compare this result to that observed in Experiment 4, we first obtained *r*s from each t-test, then compared *r*s with the *r.test* function in R which compares Fisher *r*-to-*z* transformed correlations, which revealed a significant difference, $z = 2.22$, $p = .026$, confirming our prediction that increasing cheating incentives would amplify the effect of visual disorder on cheating.

Experiment 6: The Effect of Visual Disorder on Rule-Breaking Part III

One concern with Experiments 4 and 5 was that having people rate disorder may have driven the observed cheating effects, perhaps by inadvertently causing them to think about social-disorder cues. Alternatively, we may have merely increase the salience of visual disorder as was our intention. In case of the former, we conducted Experiment 6, which was identical to

Experiment 5, except that we had people rate *preference* instead of disorder during both the training phase and the manipulation phase. Thus, there was not a single explicit mention of "order" or "disorder" in this experiment. This should alleviate any concern about having people rate disorder in Experiment 4 and 5. Because rating preference presumably would reduce the salience of visual disorder, we predicted that the effect of visual disorder on cheating would be attenuated compared to in Experiment 5.

Method

Participants and design. 394 US-based adults (202 men, 189 women, 3 other) were recruited from AMT and participated in this two-condition (visually disordered stimuli vs. visually ordered stimuli) between-subjects experiments. Sample size and stopping rule were the same as in Experiment 5. Ages ranged from 19 to 76 (M = 34.25, SD = 11.07). 327 participants identified primarily as White/Caucasian, 27 as Asian/Asian American, 20 as Black/African American, 9 as Hispanic/Latino, 4 as multiple ethnicities, 3 as Native Hawaiian, and 2 as Native American. The median experiment duration was 13 minutes and 35 seconds and participants were compensated $0.50 for participating plus a bonus of up to $2.00 (same as in Experiment 5). Informed consent was administered by the IRB of the University of Chicago.

Results

Preference ratings. An independent-samples t-test revealed that the visual disorder manipulation did not have a significant effect on preference ratings, $t(392) = 0.13$, $p = .897$, with the visually disordered stimuli (M = 3.31, SD = 0.90) receiving virtually the same preference ratings as the visually ordered stimuli (M = 3.32, SD = 1.09). Thus any effect of our manipulation on cheating could not be attributed to systematic differences in preference.

Cheating analysis: actual performance vs. reported performance. Cheating was assessed the same way as in Experiment 5. Five participants (1.3% of the sample) were excluded from the cheating analysis for performing perfectly on the Matrices Test since it would be impossible for them to cheat. Actual performance and reported performance were imperfectly correlated at $r = .52$ indicating that the procedure encouraged people to cheat. A linear mixed-model with performance on the Matrices Test predicted by visual condition, actual vs. reported, and their interaction as fixed factors and a random intercept for each participant. revealed a significant main effect of actual vs. reported, $t = 11.10$, $p < .001$, with participants across visual conditions reporting 53% higher performance ($M = 4.54$, $SD = 3.26$) than their actual performance ($M = 2.97$, $SD = 2.06$) on the Matrices Test. Importantly, there was again a significant interaction between actual vs. reported and visual condition, $t = 2.08$, $p = .038$, with participants in the visual-disorder condition reporting 63% higher performance than their actual performance and participants in the visual-order condition reporting 43% higher performance than their actual performance (see Figure 6a). The simple effect of visual condition within reported performance was also significant, $t = 2.29$, $p = .023$. A follow-up test of multivariate simple effects of actual vs. reported performance within the visual-disorder and visual-order conditions revealed that the effect size in the visual-disorder condition, $\eta_p^2 = .183$, was nearly twice as large as the effect size in the visual-order condition, $\eta_p^2 = .094$. These results corroborate that visual disorder encourages cheating, and this effect is not due to rating disorder.

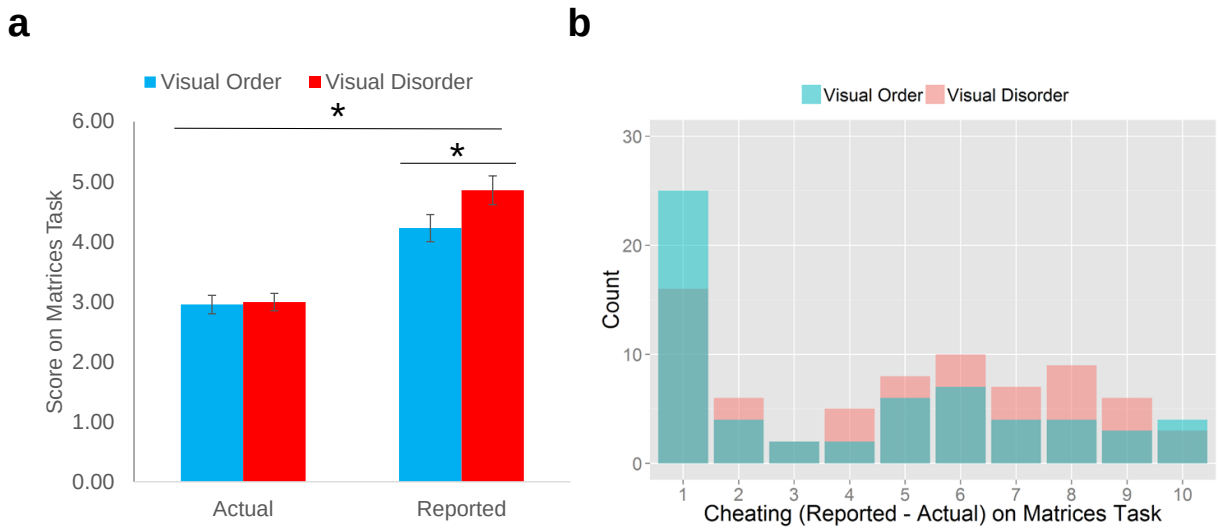


Figure 6. Visual disorder encourages cheating (Experiment 6) (a) Actual vs. reported performance by condition in Experiment 6. Error bars indicate mean ± s.e.m. (b) Magnitude of cheating (reported performance – actual performance) in Experiments 6. * $p < .05$.

Cheating analysis: likelihood of cheating. A chi-square test of independence conducted on a condition-by-cheating (yes/no) contingency table was not significant, $\chi^2(1, N = 389) = 1.29$, $p = .257$, $\phi = 0.058$, OR = 1.27, however, there was a descriptive difference in the predicted direction, with 37% of participants cheating in the visual-disorder condition and 32% of participants cheating in the visual-order condition (17% relative increase). To compare this result to that observed in Experiments 5 and 4, we computed z as before. The result of the chi-squared test in Experiment 6 did not significantly differ from the results of the chi-squared test in Experiment 5, $\delta = -0.24$, $z = -0.79$, $p = .428$, or in Experiment 4, $\delta = 0.11$, $z = -0.38$, $p = .703$.

Cheating analysis: magnitude of cheating. An independent samples t-test revealed a significant effect of visual disorder on magnitude of cheating as in Experiment 5, $t(387) = 2.08$, $p = .038$, $d = 0.21$, with those in the visual-disorder group ($M = 1.86$, $SD = 3.00$) cheating by 46% larger magnitude than those in the visual-order condition ($M = 1.27$, $SD = 2.57$) (see Figure 6b).

To compare this result to that observed in Experiments 5 and 4, we again used the *r.test* function in R. The result of the t-test in Experiment 6 did not significantly differ from the results of the t-test in Experiment 5, $z = -1.04$, $p = .297$, or in Experiment 4, $z = 1.17$, $p = .243$.

Considering the results from Experiments 4-6 together, we conclude that visual disorder is indeed sufficient to encourage rule-breaking. When cheating incentives were sufficiently large and visual disorder was salient (Experiment 5), the effect of visual disorder on cheating was largest. When the salience of visual disorder was reduced (Experiment 6), the effect of visual disorder on cheating was still marked but weaker. We made two noteworthy observations here. First, we note that the effect of visual disorder on cheating magnitude was significantly different between Experiments 5 and 4, but not between Experiments 6 and 5, suggesting that rating disorder vs. preference mattered less than increasing cheating incentives for increasing cheating magnitude (however, there were no significant differences in cheating likelihood across Experiments 4-6). Second, we note that the descriptively weaker cheating effect in Experiment 6 points to the possibility that, although visual disorder cues alone may encourage rule-breaking, there could be some top-down processes at work. However, this is not complex social reasoning of the kind put forward by broken windows theorists, rather it may have to do with priming visual disorder and its associations. This is a topic we consider in the next experiment and in the general discussion.

Experiment 7: Writing About Visually Ordered vs. Disordered Stimuli

Although Experiment 6 ruled out that rating disorder drove the cheating effects in Experiments 4 and 5, it did not directly test the possibility that rating disorder caused people to overtly think about social-disorder cues. Some may even argue that the visual disorder stimuli

themselves are imbued with social meaning (e.g., Sampson & Raudenbush, 2004). In Experiment 7, we tested these possibilities by having people freely write about the visually ordered vs. visually disordered stimuli, then analyzing their responses to see what these stimuli caused people to overtly think about. We did not expect to find any evidence that these stimuli caused overt thinking about social-disorder cues. In addition to testing this prediction, this experiment could provide some insight into the possible mechanisms driving the effects observed in Experiments 4-6.

Experimental Method

Participants and design. 98 US-based adults (51 men, 47 women) were recruited from AMT and participated in this two-condition (visually disordered stimuli vs. visually ordered stimuli) between-subjects experiments. Sample size and stopping rule were based on our goal to collect more than twice the data as collected in early thought-listing work (e.g., Petty & Cacioppo, 1977). Ages ranged from 19 to 70 (M = 34.32, SD = 10.04). 82 participants identified primarily as White/Caucasian, 8 as Asian/Asian American, 4 as Hispanic/Latino, 2 as Black/African American, 1 as multiple ethnicities, and 1 as Native American. The median experiment duration was 14 minutes and 44 seconds and participants were compensated $1.00 for participating and a bonus of up to $2.00 based on their performance in the Matrices Test (based on actual performance in this experiment vs. reported performance in Experiments 4-6, i.e., participants did not grade themselves in this task). Informed consent was administered by the IRB of the University of Chicago.

Procedure. The procedure was the same as in Experiment 5 except that instead of the self-grading task, participants were asked to do a task in which they wrote about the thoughts

they had while viewing the visually ordered vs. visually disordered stimuli. We decided to follow the same procedure as in Experiment 5 because between the two experiments in which participants rated disorder, Experiment 5 produced the larger cheating effect.

In the writing task, participants were instructed with the following:

> "What went through your mind while looking at those 30 images (i.e., the second set of images you just viewed)? Write down the thoughts you had, whatever they may be. Your responses will be completely anonymous. You have 2 minutes to write about the images, then you will automatically continue."

They were then provided with a text box in which they could write freely. After two minutes, they automatically advanced to the next page.

Preprocessing

We first ran a spellcheck to correct misspelled words that could have been counted as different terms from those which were intended. We then turned each participant's response into a document to create a document-term matrix. Utilizing the *tm* package in R, we sequentially transformed all text to lowercase, removed punctuation, stripped digits, removed standard stop words (see list in supplementary materials), stemmed all words with Porter's (1980) stemming algorithm, removed custom stop words (see list in supplementary materials), and stripped whitespace. Finally, we manually dealt with any remaining term consolidation issues we could find in the resulting terms. Only terms that were present in more than 10% of the documents in a given condition were included in the document-term matrix.

Notes on Analytic Method

For part of our analysis, we utilized the *plsRglm* R package (Bertrand, Meyer, & Maumy-Bertrand, 2014) which provides functions for partial least squares (PLS) modeling. PLS was used because of the large number of possibly correlated explanatory variables we were testing (each term is an explanatory variable). PLS fits linear models based on orthogonal linear combinations of the explanatory variables (factors) that are obtained in a way that attempts to maximize the covariance between the two sides of the equation in a dimensionality-reduced space. From there, we could estimate the weight of individual terms for predicting visual condition and their bootstrap confidence intervals.

Results and Discussion

Visual condition had marginally significant effects on how much meaningful content people wrote about as indicated both by frequency of term use, Welch's $t = 1.69$, $p = .095$, $d = .34$, and frequency of unique terms, Welch's $t = 1.86$, $p = .066$, $d = .38$, with participants in the visual-disorder condition using terms 31% more frequently ($M = 7.73$, $SD = 4.47$) and 30% more unique terms ($M = 5.92$, $SD = 2.95$) than participants in the visual-order condition ($M = 6.24$, $SD = 4.27$; $M = 4.80$, $SD = 3.01$; respectively) (see Figure 7 for term clouds depicting frequencies of term use per condition). In terms of total counts, those in the visual-disorder condition ($n = 49$) used terms 379 times and used 290 unique terms whereas those in the visual-order condition ($n = 49$) used terms 306 times and used 235 unique terms, or 24% more term use and 23% more unique terms than those in the visual-order condition. These results provide a little evidence that visual disorder increases the amount and diversity of thought, which may relate with an information-load mechanism. However, they also suggest that visual disorder did not increase

major secondary information load during the writing task, which may decrease writing length and complexity (Ransdell, Levy, & Kellogg, 2002).

Figure 7: Term clouds for the visual-order (left) and visual-disorder (right) conditions. Term size and color denote frequency of use. Only terms present in more than 10% of the documents of a given condition are presented. Any incomplete words are due to term stemming.

We next conducted the PLS analysis to test which terms significantly predicted visual condition. We decided on a six-component model (R^2 = .68) based on it having the lowest Akaike information criterion (AIC) = 46.95. The *bootpls* package in R was utilized to calculate 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals for each term coefficient. Bootstrapping was done with 1,000 bootstrap samples. Table 2 presents the coefficients and confidence intervals of terms that significantly predicted visual condition. We made some noteworthy observations. First, there were more terms that significantly predicted the visual-disorder condition than there were terms that significantly predicted the visual-order condition, and furthermore, the terms predicting the visual-disorder condition were all weighted more than the terms predicting the visual-order condition. This may be due to those in the visual-

disorder condition writing more meaningful content. We also found some expected terms such as "disord[er]" predicting the visual-disorder condition and "even" predicting the visual-order condition. Interestingly, the term "think" significantly predicted the visual-order condition whereas the term "feel" was used more frequently in the visual-disorder condition (though "feel" was not a significant predictor in the PLS model). We also noticed that the term "build[ing]" significantly predicted the visual-order condition, which may have social connotations, but we note that only 5 of 49 participants in the visual-order condition used this term. These participants may have either imagined building-like structures in the scrambled-edge stimuli that were not in the original scene images, or they detected buildings from the original scene images despite our scrambling procedure. We do not think it is a major concern because only a few people used this term and also we assume that "building" is not a social-disorder cue.

Table 2: Terms Significantly Predicting Condition in Experiment 6

Terms	Weights	95% CI Lower	95% CI Upper
See	0.37	0.27	0.65
sometim	0.33	0.18	0.57
First	0.31	0.21	0.61
Disord	0.29	0.07	0.54
Black	0.24	0.13	0.45
Find	0.21	0.01	0.42
Task	0.21	0.13	0.41
Number	0.21	0.08	0.42
Compar	-0.15	-0.39	-0.06
Build	-0.18	-0.39	-0.03
Appear	-0.20	-0.44	-0.10
One	-0.25	-0.48	-0.10
Even	-0.27	-0.51	-0.13
Think	-0.28	-0.48	-0.18

Note. Terms are ordered from most positive weight to most negative weight in the PLS model. Positive weights indicate predicting the visual-disorder condition and negative weights indicate predicting the visual-order condition. 95% CI was determined based on BCa bootstraps.

Overall, it was apparent from the term frequency list and the PLS analysis that our manipulation of visual disorder did not prime overt thinking about social-disorder cues. Rather, the terms frequently used across conditions such as “line”, “pattern”, and “symmetr[y]”, and the terms that distinguished visual condition in the PLS analysis such as “see”, “disord[er]”, “black”, “appear”, and “even”, reflect overt thinking about visual cues. Assuming that several of these terms have semantic connotations, the results of this study suggest that systematic differences in semantic associations may play a role in the cheating effect. Furthermore, the term-frequency analysis provided evidence, albeit weak, that systematic differences in information load may (also) play a role in the cheating effect.

Chapter 1 Discussion

This study set out to answer two major questions. First, what are some of the key visual features that define visual disorder? Second, are these visual disorder cues alone sufficient to encourage rule-breaking despite the absence of social disorder cues? Our first set of experiments (Experiments 1-3) showed that non-straight edge density and asymmetry are key components of visual disorder. More generally, these experiments suggest that spatial features are more important for visual disorder than are color features. Such insights into the building blocks of visual disorder are important if we are to make significant advancements in our understanding of phenomena relevant to broken windows theory and more broadly how visual stimulation/processing can affect non-visual behavior. Taking on the challenge of quantifying elements of perceived disorder is warranted by the demonstrated societal impact of this theory. Our second set of experiments (Experiments 4-6) demonstrated that exposure to visual disorder

cues alone could encourage rule-breaking behavior. The final experiment (Experiment 7) was a beginning exploration regarding the mechanisms driving the effects of visual disorder on cheating behavior by examining what people overtly thought about when viewing visually ordered vs. visually disordered stimuli. The broad implication of this study is that established theories of rule-breaking that assume that *reasoning about social cues* is necessary, should be reconsidered (e.g., Kelling & Coles, 1997; Sampson & Raudenbush, 2004).

To elaborate on Experiments 1-3, although they start to answer the broader question of what makes an environment visually disordered, they surely do not exhaustively answer this question. It would be interesting to look into other possible visual disorder cues such as variability in edge orientation. In particular, although curved edges were seen as disorderly in our experiments, intuition says that several curved edges arranged in parallel to one another would be more orderly than curved edges arranged in a haphazard way. Our results linking symmetry and order support this intuition but not directly. We encourage researchers to derive other metrics to quantify and manipulate possible visual disorder cues to further our understanding of what makes environments visually disordered.

To elaborate on Experiments 4-6, the results suggest that the "cheating effect" of visual disorder actually has two separable components. Visual disorder not only increased the amount by which cheaters cheat (*cheating magnitude*), it also encouraged people who normally would not cheat to cheat (*cheating likelihood*). That is, cheaters were nudged towards cheating more and noncheaters were nudged towards cheating at all. A meta-analysis of Experiments 4-6 revealed highly significant effects of our manipulation of visual disorder on both cheating likelihood, $\chi^2(1, N = 1{,}189) = 8.87$, $p = .003$, $\phi = 0.086$, OR = 1.43, and cheating magnitude,

$t(1,187) = 3.47$, $p < .001$, $d = 0.20$, with visual disorder increasing cheating likelihood by 25% on average and cheating magnitude by 42% on average. This sort of cheating could have major economic and societal consequences. Imagine if the amount by which people underreported their taxes increased by just 1%—billions of dollars would be lost.

One of the big questions remaining is, why is this happening? Although Experiment 7 started to touch on this issue, it would take more research to nail down the mechanisms involved in why visual disorder is causing cheating. A careful examination of specific mechanisms is beyond the scope of this chapter. That said, the results of Experiment 7 point to two classes of mechanisms that we can speculate about—the first reflecting an information processing approach and the second reflecting a priming or spreading activation approach. First, the term-frequency analysis suggests that visual disorder cues may be more informationally burdensome than the visual order cues, at least in the sense of increasing amount and diversity of thought. This may relate with visually disordered stimuli being less redundant (i.e., fewer spatially predictable patterns) and conveying more information than visually ordered stimuli.[3] These aspects of visual disorder may make viewing visual disorder less perceptually fluent than viewing visual order (see Field, 1987; Kinchla, 1977; Olshausen & Field, 1996; Witkin & Tenenbaum, 1983). If either of these possibilities are true—that visual disorder is more informationally burdensome or that visual disorder is less perceptually fluent—then there could be a whole slew of downstream consequences on judgments and even complex behaviors. Regarding information "overload", this could fatigue cognitive resources necessary for self-regulation (Hofmann, Friese,

[3] As a side note, judgments of disorder may themselves be guided by encoding difficulty to the extent that they are related to judgments of randomness (Falk & Konold, 1997).

Schmeichel, & Baddeley, 2011; Kaplan & Berman, 2010; Kotabe & Hofmann, 2015). Regarding perceptual disfluency, this could affect judgments such as of familiarity that are involved in the production of complex behaviors (Alter & Oppenheimer, 2009).

As for the second class of mechanisms, we speculate that visual disorder could have a priming or spreading activation effect that is involved in the production of complex behaviors. For example, the lack of "symmetry" and "patterns" (both terms used frequently by participants in Experiment 7) may give rise to a mindset that things are random and uncontrollable, which may reduce the motivation for self-control (Kotabe, 2014; see also Tullett et al., 2015). Or, they may activate mental metaphors which are manifested in a family of linguistic metaphors relevant to rule-breaking such as in "he's as *straight* as an arrow" and "he's *bending* the rules" (for a review of research on the linkage between spatial representations and abstract concepts, see Casasanto & Bottini, 2014). There is a rich literature on feedforward and feedback projections from visual cortex (e.g., V1) to higher cortical areas, thus providing physiological evidence that low-level visual processing may interact with higher cortical areas involved in the production of complex behaviors (e.g., Felleman & Essen, 1991; Gilbert & Li, 2013; Lamme & Roelfsema, 2000; McIntosh et al., 1994), though higher cortical areas may not need to be involved at least in the interpretation of simple visual information (Brascamp et al., 2015). It is also possible that mechanisms from both classes may be at work and could interact with each other. Regardless, these possible mechanisms paint a completely different picture from the dominant explanations for broken windows phenomena. As such, they point to a vast and unattended area of research, which we encourage researchers to venture into.

Another important remaining question concerns what we mean, specifically, when we say that visual disorder "encourages" cheating. According to integrative self-control theory (Kotabe & Hofmann, 2015), there are several components at work in situations involving competing desires and goals (e.g., a desire to cheat vs. a goal to be honest in this case), each of which can increase or decrease the likelihood of one of these motivations being enacted. For example, as alluded to in the previous paragraph, viewing visual disorder could fatigue cognitive-control capacity via information "overload", it could decrease cognitive-control motivation via changing mindsets, or it could increase the desire to cheat via some priming or spreading activation process. And to complicate things, these are likely not independent processes. For example, increasing desire strength may activate or inhibit control goals (Fishbach, Friedman, & Kruglanski, 2003) and fatiguing control capacity may increase desire strength (K. D. Vohs et al., 2013; Wagner, Altman, Boswell, Kelley, & Heatherton, 2013). Teasing apart these processes is a major challenge (Kotabe & Hofmann, in press), and thus much work remains.

To conclude, research on environmental disorder has tended to focus on its consequences (e.g., Braga & Bond, 2008; Keizer et al., 2008; Kelling & Coles, 1997), yet little is known about what makes an environment disorderly in the first place. As this work demonstrates, deconstructing the features of disorderly and orderly environments can help us to understand how disorderly environments affect us in ways harmful to ourselves and to society. In addition, this approach could inform the design of environments—both real and virtual. Considering the observed effect of visual disorder on rule-breaking behavior, and the evidence that rule-breaking behaviors spread (Keizer et al., 2008), we should take (imparting) visual disorder in our environments seriously.

CHAPTER 2: THE NATURE-DISORDER PARADOX

Exposure to natural environments has been shown to be beneficial for humans, whereas exposure to disorderly environments has been shown to be detrimental (Chae & Zhu, 2014; Geis & Ross, 1998; Heintzelman, Trent, & King, 2013; Keizer et al., 2008; Kotabe, 2014; Perkins & Taylor, 1996; Ross, 2000; Tullett et al., 2015; Kathleen D. Vohs et al., 2013; J. Q. Wilson & Kelling, 1982). For example, 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., 2012; Berman, Jonides, & Kaplan, 2008), increase positive affect (Berman et al., 2012), decrease negative affect (Bratman, Daily, Levy, & Gross, 2015; Bratman, Hamilton, Hahn, Daily, & Gross, 2015), decrease aggression (Kuo & Sullivan, 2001), and decrease crime (Kuo & Sullivan, 2001). Exposure to disorderly environments, in contrast, may encourage rule-breaking and criminal behavior (see Chapter 1; see also Keizer et al., 2008), worsen self-control and cognitive-control (Chae & Zhu, 2014), decrease a sense of meaning in life (Heintzelman et al., 2013), and increase negative affect (Ross, 2000; Tullett et al., 2015).

Countless studies have shown that natural environments are aesthetically preferred over built environments. We define "aesthetic preference" as a like-dislike affective response (Zajonc, 1980) elicited by visual exposure to scenes (as did Ulrich, 1983). It may be distinct from other components of reward such as 'wanting' and 'learning' which come before and after, respectively (Berridge, Robinson, & Aldridge, 2009). The popular biophilia hypothesis proposes that humans have an affinity to the natural and the living that is rooted in our evolutionary history (E. O. Wilson, 1984). This strong aesthetic preference for natural environments has been theoretically and empirically linked to nature's restorative potential (Han, 2010; Hartig & Staats,

2006; Purcell, Peron, & Berto, 2001; Staats, Van Gemerden, & Hartig, 2010; Ulrich, 1983; Van den Berg, Koole, & van der Wulp, 2003). But, paradoxically, nature *is judged* as being disorderly. We have found repeatedly across multiple datasets that naturalness and disorder are significantly correlated (r = [.35, .42]). How is it that nature scenes are aesthetically preferred when they are relatively disorderly? To add to the confusion, highly natural and highly disorderly scenes seem to share some similar low-level visual features, such as broken and non-straight edges and asymmetry (Berman et al., 2014; Kotabe, Kardan, & Berman, 2016b). How is it that at the basic visual level these dimensions are not clearly separable, but in terms of psychological effects, they could not be much more different?

Making sense of this paradox would be useful for psychological theories concerning the joint influence of lower-level and higher-level perceptual inputs on affect and cognition. There is little work that has systematically separated the low- and high-level inputs of environmental scenes, much less whether there are differential effects of the low- vs. high-level inputs vs. their interaction on important psychological variables (Kardan, Demiralp, et al., 2015; but see Kardan et al., in press). At a high-level, if disorder affects aesthetic preferences for natural environments, it would provide evidence against the idea that "nature" is a unitary construct that uniformly affects people. If naturalness and disorder relate with each other and with aesthetic preference to different degrees when only low-level visual features are present (i.e., when semantics are obscured) compared to when high-level scene semantics[1] are preserved, it would suggest that the presence of recognizable entities in an environment can suppress or strengthen the relationships

[1] We use the term "semantics" to refer to concepts and categories associated with recognizable entities in the environment.

between naturalness, disorder, and aesthetic preference. It would support our view that naturalness and its aesthetics are complex and nuanced, involving various lower-level and higher-level inputs (Berman et al., 2014; Kardan, Demiralp, et al., 2015). Simply finding that disorder affects aesthetic preference for a scene, independently from naturalness, would answer the following question: does disorder matter in nature? Surprisingly little is known about this because virtually all of the research on environmental disorder has been conducted in built environments.

There are at least three possible explanations for the nature-disorder paradox: (a) The positive effect of naturalness on aesthetic preference trumps the negative effect of disorder (*nature-trumps-disorder hypothesis*). That is, people independently like naturalness and dislike disorder, but the effect of naturalness is stronger than the effect of disorder. (b) Disorder does not affect aesthetic preference in natural environments. That is, people dislike disorder in built environments but it does not bother them in natural environments (e.g., because people *expect* nature to be disorderly) (*harmless-disorder hypothesis*). (c) Disorder has a positive effect on aesthetic preference in natural environments (*beneficial-disorder hypothesis*). That is, people dislike disorder in built environments but they actually like when natural environments are disorderly (e.g., because it is more wild and reminiscent of their ancestral experience, consistent with an evolutionary perspective). All of these hypotheses assume that scene semantics play a key role. Through a series of experiment, we rule in favor of the nature-trumps-disorder hypothesis and rule out the harmless-disorder and beneficial-disorder hypotheses. Furthermore, we show that when scene semantics are removed, the nature-trumps-disorder effect disappears, demonstrating that the nature-trumps-disorder effect is driven by higher-level semantics.

General Method

We sampled broadly from real-world environments by using diverse sets of images of environmental scenes as well as words that varied in terms of naturalness and disorder (see Figure 8 for examples; all images utilized in this study can be downloaded here in original resolution: goo.gl/za9seG). One set was the set of 260 scene images we used in the study presented in Chapter 1. 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 images contained not only trees, parks, etc. but also waves, mountains, and lava). In Experiments 1a-c and 2a-c, we used the scene images in their original form. In Experiments 3a-f, we extracted and scrambled visual features from the scene images to address the role of scene semantics. For each image—including the original images and the manipulated versions—we calculated mean ratings of naturalness, disorder, and aesthetic preference. Data analysis was conducted at the level of individual image-level means. For Experiments 4a-c the method was similar but instead of images we used words as stimuli to further investigate the role of semantics vs. visual features.

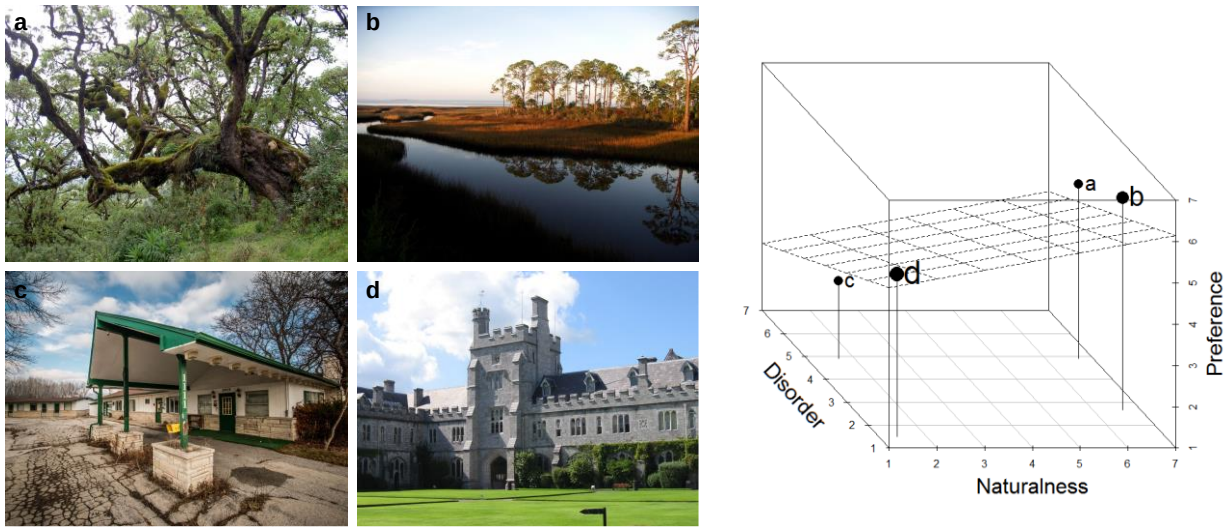


Figure 8: Chapter 2 example stimuli. On the left, four scenes from the set of 916 scene images used in Experiments 2a-c that exemplify the coexistence of **(**a) naturalness and disorder; (b) naturalness and order; (c) builtness and disorder; and (d) builtness and order. On the right, these scenes are mapped in three-dimensional space relative to the plane predicted when regressing aesthetic preferences on naturalness and disorder in this dataset.

Experiments 1a-c: Reanalyzing Previously Collected Data

We analyzed previously collected naturalness (Experiment 1a), disorder (Experiment 1b), and aesthetic preference (Experiment 1c) ratings for 260 (Kardan, Demiralp, et al., 2015; Kotabe et al., 2016b) environmental scenes as a first test of the three competing hypotheses. We also quantified spatial and color visual features as in (Berman et al., 2014; Kardan, Demiralp, et al., 2015) to statistically control for low-level visual variation in the environmental scenes. This way we could test whether the relative effects of nature and disorder on aesthetic preference depend on low-level visual features, or if semantics alone are primarily at work.

Results

Per the nature-disorder paradox, naturalness and disorder were significantly correlated at $r = .35$, $p < .001$ (see Table 3). Naturalness was significantly correlated with aesthetic preference

at $r = .73$, $p < .001$ but disorder was not significantly correlated with aesthetic preference, $r = .08$, $p = .177$. After controlling for disorder, naturalness was partially correlated with aesthetic preference at $r_p = .81$, $p < .001$ and, after controlling for naturalness, disorder was partially correlated with aesthetic preference at $r_p = -.52$, $p < .001$. The relative changes in these correlations indicates that builtness-naturalness suppressed the association between order-disorder and aesthetic preference more than order-disorder suppressed the association between builtness-naturalness and aesthetic preference, consistent with the nature-trumps-disorder hypothesis.

Table 3: Correlations Between Naturalness, Disorder, and Preference Ratings Across all Reported Experiments

	260 scenes (Experiments 1a-c)			916 scenes (Experiments 2a-c)		
	Naturalness	Disorder	Aesthetic Preference	Naturalness	Disorder	Aesthetic Preference
Naturalness	–			–		
Disorder	.35***	–		.36***	–	
Aesthetic Preference	.73***	-.08	–	.46***	-.16***	–
	260 scrambled-edge stimuli (Experiment 3a-c)			260 scrambled-color stimuli (Experiment 3d-f)		
Naturalness	–			–		
Disorder	.01	–		-.31***	–	
Aesthetic Preference	.00	-.64***	–	.02	-.36***	–
	Experiments 4a-c (632 words)					
Naturalness	–					
Disorder	.37***	–				
Noun Preference	.34***	-.22***	–			
*** p < 0.001						

Next, we simultaneously regressed aesthetic preference on naturalness, disorder, and their interaction (see Table 4, Experiments 1a-c, Model 1). These factors explained almost two thirds of the variance in aesthetic preferences, $R^2_{adj} = .65$. Both naturalness ($\eta_p^2 = .65$) and disorder ($\eta_p^2 = .28$) significantly predicted aesthetic preferences. A linear contrast indicated that the effect of

perceived naturalness on aesthetic preference was significantly larger than the effect of perceived disorder, $F(1, 256) = 117.17$, $p < .001$, further supporting the nature-trumps-disorder hypothesis. In addition, we calculated the relative importance of naturalness and disorder for predicting aesthetic preference with the *relaimpo* R package (Grömping, 2006), which takes into account intercorrelations between variables. Across all eight metrics calculated by the package, naturalness was estimated to be more important than disorder for aesthetic preference—e.g., the recommended *lmg* method (Lindeman, Merenda, & Gold, 1980), which partitions R^2 by averaging over orders, estimated that 90% of the variance in the model was explained by naturalness vs. 10% by disorder. Regarding the harmless-disorder and beneficial-disorder hypotheses, there was a marginal *negative* interaction between naturalness and disorder ($\eta_p^2 = .01$) suggesting, if anything, that disorder may actually have a stronger negative effect in natural environments than in built environments, contradicting the harmless-disorder and beneficial-disorder hypotheses.

Table 4: Regression Models in Experiments 1a-c and 2a-c

		260 scenes (Experiment 1a-c)		916 scenes (Experiment 2a-c)	
		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 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. Aesthetic preferences regressed on naturalness, disorder, and their interaction, with and without controlling for low-level visual features, in Experiments 1a-c and Experiments 2a-c. Standardized coefficients not in parentheses and standard errors in parentheses.

To examine the independent high-level effects (e.g., semantics) of naturalness and disorder, we statistically controlled for low-level color and spatial factors (i.e., low-level visual features) in another regression model (see Table 4, Experiment 1a-c, Model 2). Visual features were quantified as in (Berman et al., 2014; Kardan, Demiralp, et al., 2015) (see Table 4 for a list

of the features).[2] Both naturalness ($\eta_p^2 = .51$) and disorder ($\eta_p^2 = .28$) still significantly predicted aesthetic preferences, 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 model was explained by naturalness vs. 9% by disorder. Furthermore, the significant negative interaction between naturalness and disorder, $\beta = -0.09$, $t(246) = -2.12$, $p = .035$, $\eta_p^2 = .02$, contradicts the harmless-disorder and beneficial-disorder hypotheses. By controlling for variance in low-level features, these results provide first evidence that semantics play an important role in driving the nature-trumps-disorder effect. The following experiments further test this possibility.

Experiment 2a-c: Replicating With New Scene Images

These experiments tested whether the results from Experiments 1a-c would replicate with a larger and more semantically diverse set of images.

Method

Participants and design. 702 US-based adults (392 women, 308 men, 2 other) were recruited from AMT and were randomly assigned to one of the three sub-experiments. Sample size was determined by our goal to receive ~20 ratings per image. 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

[2] One difference was that we improved how hue and SD hue were calculated. Because the hue of a pixel is an angular value, mean and standard deviation of hue of the images were calculated using circular mean and standard deviation (Circular Statistics Toolbox for MATLAB).

American, 39 as Asian/Asian American, 37 as Hispanic/Latino, 8 as "multiple ethnicities," 5 as Native American/Alaska Native, and 3 as "other." The median experiment duration was 9 minutes and 39 seconds and participants were compensated $1.00 for participating. Informed consent was administered by the Institutional Review Board (IRB) of the University of Chicago.

Materials. Images were selected from the SUN image database (Xiao et al., 2010); a database that contained a more semantically diverse set of images than used previously (e.g., including scenes of open sky, waves, and volcanoes). We sampled more orderly nature scenes and more disorderly built scenes as those categories were less sampled in Experiments 1a-c. As in Experiments 1a-c, only scenes without human or animal presence were selected. This yielded a set of 1,105 images in total.

Procedure. Participants were first given a brief introduction to the image-rating task. Then they were randomly presented 100 of the 1105 scene images on a plain white background. This randomization scheme dealt with an issue in Experiments 1b-c. In those previous experiments participants received 10 images from each quintile of naturalness when making disorder or preference ratings. Thus, each participant received the full sample of built to natural scene images, but possibly did not receive the full sample of disorderly to orderly scene images. It is possible that this randomization scheme "favored" naturalness over disorder, thus giving rise to the nature-trumps-disorder effect. The randomization scheme in Experiments 2a-c did not have this bias since images were not partitioned prior to randomization.

Regarding the image rating task, 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?" It is thought that simple like-dislike ratings reliably reflect affective discriminations (Zajonc, 1980). Participants made ratings using seven-point semantic differential scales ("very manmade" to "very natural"; "very disorderly" to "very orderly"; "strongly dislike" to "strongly like"). In addition, participants did a fourth version of this experiment in which they rated "rule-breaking" which is a semantically rich construct beyond the scope of this study (though pertinent to Chapter 1), because here our focus is on physical disorder rather than social disorder. 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, leaving 916 images for analysis. The rule-breaking data has not yet been analyzed but will be for in future research.

Results

Naturalness and disorder were again significantly correlated, $r = .36$, $p < .001$ (see Table 3). Naturalness was significantly correlated with aesthetic preference at $r = .46$, $< .001$ and disorder was significantly correlated with aesthetic preference at $r = -.16$, $p < .001$. After controlling for disorder, naturalness was partially correlated with aesthetic preference at $r_p = .56$, $p < .001$ and, after controlling for naturalness, disorder was partially correlated with aesthetic preference at $r_p = -.40$, $p < .001$. The relative changes in these correlations indicates that, in this new sample, builtness-naturalness again suppressed the association between order-disorder and aesthetic preference more than order-disorder suppressed the association between builtness-naturalness and aesthetic preference. These results are again, consistent with the nature-trumps-disorder hypothesis.

As before, we simultaneously regressed aesthetic preference on naturalness, disorder, and their interaction (see Table 4, Experiment 2a-c, Model 1). These factors explained about a third of the variance in aesthetic preference, $R^2_{adj} = .33$. This is about half the variance explained by these factors in Experiment 1a-c, supporting that the new sample of images was more semantically diverse. Both naturalness ($\eta_p^2 = .32$) and disorder ($\eta_p^2 = .15$) again significantly predicted aesthetic preference. A linear contrast indicated that the effect of naturalness on aesthetic preference was significantly larger than the effect of perceived disorder, $F(1, 912) = 43.01$, $p < .001$, supporting the nature-trumps-disorder hypothesis. Furthermore, we again calculated relative importance of naturalness and disorder for predicting aesthetic preference. Across all eight metrics calculated by the *relaimpo* package, naturalness was estimated to be more important than disorder for aesthetic preference—e.g., the recommended *lmg* method estimated that 77% of the variance in the model was explained by naturalness vs. 23% by disorder. These results suggest that even when scene semantics are substantially diversified, naturalness still trumps disorder in driving aesthetic preference. Regarding the alternative hypotheses, there was no significant interaction between naturalness and disorder, again in contradiction with the harmless-disorder and beneficial-disorder hypotheses.

Controlling for low-level color and spatial factors in another simultaneous regression model, both naturalness ($\eta_p^2 = .22$) and disorder ($\eta_p^2 = .17$) still significantly predicted environmental aesthetic preferences (see Table 4, Experiment 2a-c, Model 2). Further, a linear contrast again indicated that the effect of naturalness on aesthetic preference was significantly larger than the effect of disorder, $F(1, 902) = 18.57$, $p < .001$. Regarding relative importance, the *lmg* method estimated that 41% of the variance in the model was explained by naturalness vs.

20% by disorder. Again, there was no interaction between perceived naturalness and perceived disorder.

If naturalness and disorder independently affect aesthetic preference, it implies that there should be a strong aesthetic preference for highly ordered nature scenes (e.g., imagine a Japanese garden), as opposed to highly disordered built environments. Supporting this prediction, in both Experiments 1a and 1b, the most ordered natural scenes were most preferred and the most disordered built scenes were least preferred, with ordered and built scenes and disordered and natural scenes in between (see Figure 9).

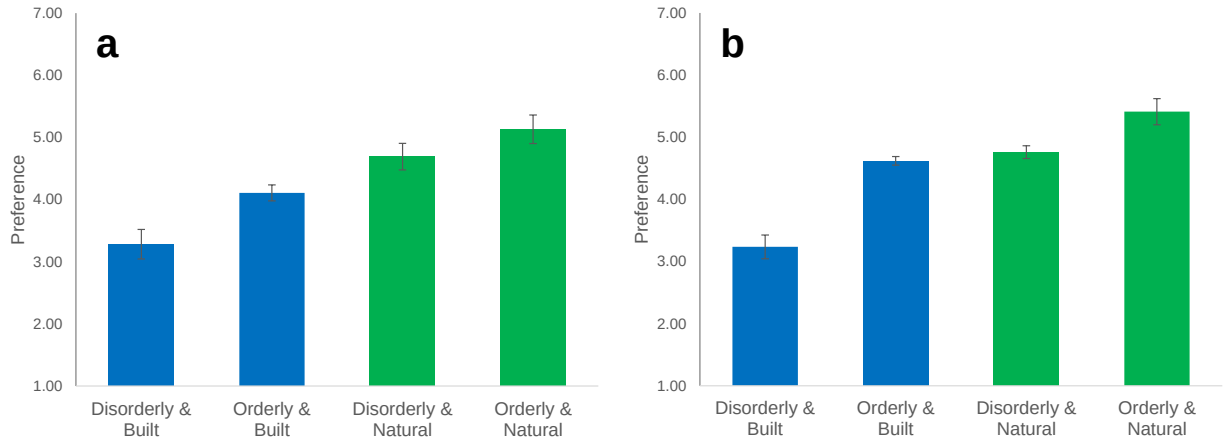


Figure 9: Mean aesthetic preference ratings. Mean aesthetic preference ratings for scene images rated in the top quintiles of builtness/naturalness and order/disorder in Experiments 1a-c (panel A) and Experiments 2a-c (panel B). Error bars indicate mean±s.e.m.

Experiment 3a-c: At the Level of Edges

Experiments 1a-c and 2a-c strongly supported that, at the level of scenes, the nature-trumps-disorder hypothesis is valid whereas the harmless-disorder and beneficial-disorder hypotheses were not supported. Statistically controlling for variations in low-level visual features

had relatively little effect, suggesting that the nature-trumps-disorder effect may not operate at the level of basic visual features. To test this possibility more rigorously, we would need to experimentally control for scene semantics. In Experiments 3a-c, we removed scene semantics by extracting and scrambling the edge features from the scene images as explained in Chapter 1. We had people rate the edge features alone in terms of naturalness (Experiment 3a), disorder (Experiment 3b), or aesthetic preference (Experiment 3c).

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. The median experiment duration was 4 minutes and 14 seconds and participants were compensated $0.50 for participating. Informed consent was administered by the IRB of the University of Chicago.

Procedure. We experimentally manipulated the images by extracting and scrambling the low-level edge features which effectively removed scene semantics (e.g., trees, buildings) (see Figure 10 and Chapter 1). Individuals rated these resulting scrambled-edge stimuli in terms of naturalness, disorder, and aesthetic preference. Thus, we could statistically estimate the degree to which naturalness and disorder predict aesthetic preference at the basic visual level of edges.

Participants were randomly presented 50 of the 260 scrambled-edge stimuli. As in Experiments 1b-c, 10 scene images were selected from each of the original images' naturalness

quintiles. In this case, this randomization scheme, if anything, would make this experiment extra conservative by favoring naturalness over disorder. That is, if the randomization scheme were biased in this way, we would be *less* likely to confirm our prediction that the nature-trumps-disorder effect goes away after removing scene semantics. For each image, participants rated either naturalness (Experiment 3a), disorder (Experiment 3b), or aesthetic preference (Experiment 3c) using seven-level semantic differential scales as before.

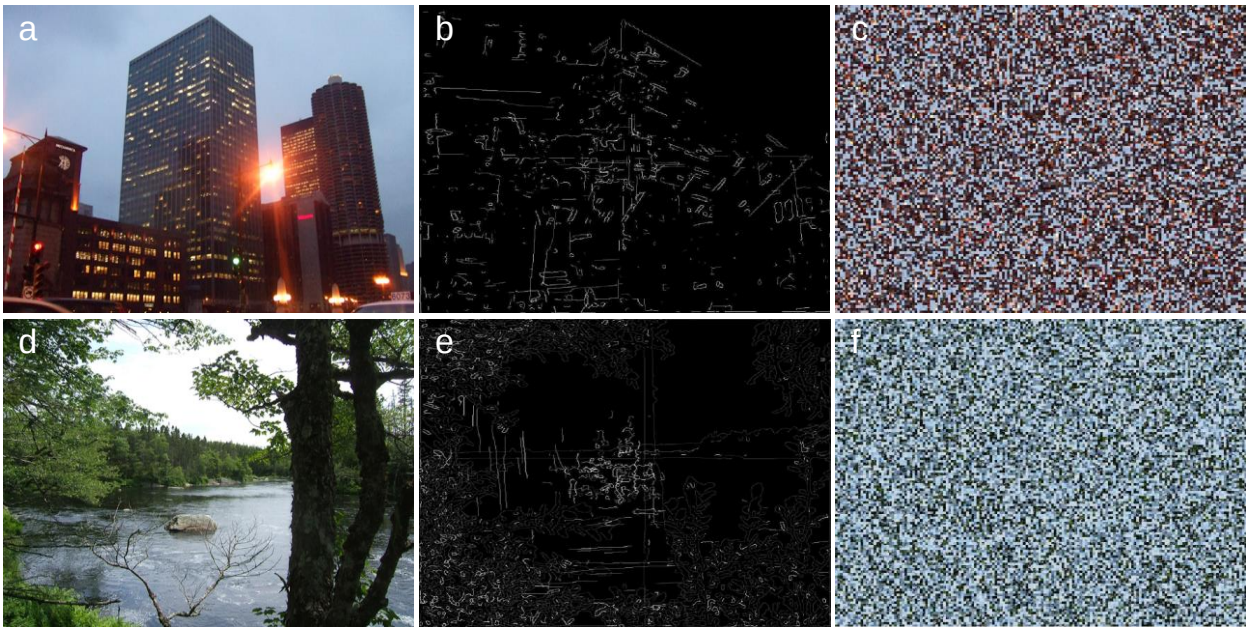


Figure 10: Examples of the highest-rated built and highest-rated natural scene images from the set of 260 scene images and their derived stimuli. (a) Original highly-built scene image (Experiment 1a-c); (b) its derived scrambled- edge stimulus (Experiment 3a-c); and (c) its scrambled-color stimulus (Experiment 3d-f). (d) Original highly-natural scene image (Experiment 1a-c); (e) its derived scrambled-edge stimulus (Experiment 3a-c); and (f) its scrambled-color stimulus (Experiment 3d-f).

Results

Naturalness and disorder were not significantly correlated for these scrambled-edge stimuli (see Table 3), suggesting that the nature-disorder paradox does not occur at the level of edges. Naturalness was also not significantly correlated with aesthetic preference but disorder was at r = -.64, p < .001. After controlling for disorder, naturalness was still not correlated with aesthetic preference and, after controlling for naturalness, disorder was still partially correlated with aesthetic preference at r_p = -.64, < .001. The absence of significant correlations with naturalness suggests that naturalness may have a weak presence at the level of edges, whereas the strong (partial) correlation of disorder with aesthetic preference (similar to the partial correlation in Experiments 1a-c, r_p = -.52) suggests that disorder is preserved at the level of edges (without semantics). Consistent with this possibility, naturalness did not suppress the association between order-disorder and aesthetic preference, providing first evidence that the nature-trumps-disorder effect goes away when semantics are removed.

To test this another way, we simultaneously regressed aesthetic preference on naturalness, disorder, and their interaction (R^2_{adj} = .41). Disorder, β = -0.64, t(256) = -13.10, p < .001, η_p^2 = .41, significantly predicted aesthetic preference but naturalness did not, β = 0.01, t(256) = 0.16, p = .877, η_p^2 = .00. A linear contrast indicated that the effect of perceived disorder on aesthetic preference was significantly larger than the effect of perceived naturalness, F(1, 256) = 88.94, p < .001. Furthermore, we calculated the relative importance of naturalness and disorder for predicting aesthetic preference as before. Across all eight metrics calculated, disorder was estimated to be more important than naturalness for aesthetic preference—e.g., the recommended *lmg* method estimated that >99% of the variance in the model was explained by

disorder vs. <1% by naturalness. Regarding the alternative hypotheses, the interaction was not significant, $\beta = 0.02$, $t(256) = 0.46$, $p = .649$, $\eta_p^2 = .00$. These results not only are inconsistent with the nature-trumps-disorder hypothesis, but they are exactly the opposite. That is, at the level of edges, disorder is much more important for aesthetic preference than is naturalness, which seems to have a weak presence at this level. It is possible, however, that the nature-trumps-disorder effect went away not because semantics were removed, but rather because colors were removed as well when we isolated edges. To address this possible confound, we isolated colors in the next set of experiments.

Experiment 3d-f: At the Level of Colors

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. The median experiment duration was 4 minutes and 10 seconds and participants were compensated $0.50 for participating. Informed consent was administered by the IRB of the University of Chicago.

Procedure. The procedure was the same as in Experiments 2a-c except that participants were presented the scrambled-color stimuli instead of the scrambled-edge stimuli. See Chapter 1 for full details on the color extraction and scrambling method.

Results

For the scrambled-color stimuli, naturalness and disorder were *negatively* correlated at $r = -.31$, $p < .001$ (see Table 3), suggesting that the nature-disorder paradox does not occur at the level of colors either, and, in addition, that natural colors are associated with order—a paradoxical result in and of itself that requires further research. Naturalness was again not significantly correlated with aesthetic preference but disorder was at $r = -.36$, $p < .001$. After controlling for disorder, naturalness was still not significantly correlated with aesthetic preference and, after controlling for naturalness, disorder was partially correlated with aesthetic preference at virtually the same level as before, $r_p = -.37$, $p < .001$. Although naturalness did not significantly correlate with aesthetic preference, it did significantly correlate with disorder, suggesting that naturalness may have a stronger presence at the level of colors than it does at the level of edges. Nevertheless, the minimal changes in the partial correlations indicates that, at the level of colors, builtness-naturalness had virtually no suppression effect on the association between order-disorder and aesthetic preference, providing more evidence that the nature-trumps-disorder effect disappears when semantics are removed.

Next, we simultaneously regressed aesthetic preference on naturalness, disorder, and their interaction ($R^2_{adj} = .13$). Again, disorder, $\beta = -0.39$, $t(256) = -6.41$, $p < .001$, $\eta_p^2 = .14$, but not naturalness, $\beta = -0.08$, $t(256) = -1.20$, $p = .230$, $\eta_p^2 = .01$, significantly predicted aesthetic preferences. The interaction was not significant, $\beta = 0.05$, $t(256) = 0.91$, $p = .364$, $\eta_p^2 = .00$. A linear contrast indicated that the effect of perceived disorder on aesthetic preference was significantly larger than the effect of perceived naturalness, $F(1, 256) = 22.36$, $p < .001$. We also calculated relative importance of naturalness and disorder for predicting aesthetic preference as

before. Across all eight metrics calculated, disorder was estimated to be more important than naturalness for aesthetic preference—e.g., the recommended *lmg* method estimated that 97% of the variance in the model was explained by disorder vs. 3% by naturalness. Regarding the alternative hypotheses, there was no significant interaction between naturalness and disorder, again in contradiction with the harmless-disorder and beneficial-disorder hypotheses. Like at the level of edges, at the level of colors, disorder seems to trump naturalness in terms of determining aesthetic preference. It should be noted that at the color-level, these factors explained significantly less variance in aesthetic preference (R^2_{adj} = .14, bootstrapped 95% CI_{BCa} [.05, .22]) than they did at the edge-level (R^2_{adj} = .41, bootstrapped 95% CI_{BCa} [.31, .49]), consistent with our previous work that showed that visual disorder is more a function of edges than colors (Chapter 1).

Collectively, the experiments reported so far suggest that naturalness trumps disorder in driving aesthetic preferences at the level of scenes but not at the level of basic spatial and color visual features. At the basic visual level—especially at the level of edges—disorder seems to trump naturalness in guiding aesthetic preferences. This may be due to naturalness having a weaker presence at this level than disorder. We conclude that semantics are *necessary* for the nature-trumps-disorder effect. Next, we ask, are they *sufficient*?

Experiments 4a-c: At the Level of Semantics

We tested whether semantics alone are sufficient for the nature-trumps-disorder effect. In this way, these experiments are the counterpart to the edge and color experiments. We presented people with a wide variety of word stimuli that were all nouns that ranged from more nature-related to more urban-related. Nouns conveyed the semantic information without the overt visual

information of scenes. People rated these words either in terms of naturalness (Experiment 4a), disorder (Experiment 4b), or preference (Experiment 4c).

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 word. 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. The median experiment duration was 7 minutes and 0 seconds and participants were compensated $0.50 for participating. Informed consent was administered by the IRB of the University of Chicago.

Materials. In total, 632 words, all nouns, were selected from the MRC Psycholinguistic Database (Coltheart, 1981) by a hypothesis-blind research assistant. She was instructed to select nouns that she judged as nature-related, urban-related, or neither nature- nor urban-related. In total, she classified 213 words as more nature-related (e.g., mountain, woodland), 222 words as neither nature- nor urban-related (e.g., compass, barrel), and 197 words as more urban-related (e.g., traffic, office) (see supplementary materials for the full word list).

Procedure. The 632 words were split into ten quantiles based on their Thorndike-Lorge written frequency (TL-FRQ) measure (Thorndike & Lorge, 1944). The 10 quantiles of words 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 attention check was used because we thought rating words may be boring compared to rating scene images. Participants were randomly

presented 10 words (or 9 words and an attention check item) from each randomly presented quantile, thus each participant rated 81 to 100 words that ranged widely in terms of written frequency. Participants rated naturalness (Experiment 4a), disorder (Experiment 4b), or preference (Experiment 4c) using seven-level semantic differential scales.

Results

One word ("year") was excluded from the analysis because of unusually high leverage (centered leverage value = .33) in the multiple regression reported below. All other centered leverage values were < .12.

Overall, the research assistant's judgments mapped onto mean naturalness ratings provided by the participants. Words categorized as more urban, neither urban nor natural, and more natural received M = 2.52, SD = 0.51; M = 4.75, SD = 1.06; and M = 5.88, SD = 0.61 on the naturalness scale, respectively.

For the following correlation and regression analyses, we statistically controlled for two factors for which we had data for all of the words (TL-FRQ and word length). Naturalness and disorder ratings for words were correlated to a similar degree as in the Experiments 1a-c and 2a-c in which our stimuli were scene images (see Table 3), suggesting that the nature scenes we used were rich in semantic content. Naturalness was correlated with noun preferences at r = .34 and disorder was correlated with noun preferences at r = -.22. After controlling for disorder, naturalness was partially correlated with noun preferences at r_p = .46 and, after controlling for naturalness, disorder was partially correlated with noun preferences at r_p = -.39. The relative changes in these correlations indicates that, at the level of nouns, builtness-naturalness suppressed the association between order-disorder and noun preferences to a similar degree as

order-disorder suppressed the association between builtness-naturalness and noun preferences. In fact, the latter suppression effect was slightly larger than the former. It is possible that visual features must interact with nature semantics to drive a strong suppression effect of builtness-naturalness. Next, we tested the competing hypotheses with multiple regression.

We simultaneously regressed noun preferences on naturalness, disorder, and their interaction. This model explained over a quarter of the variance in noun preferences, $R^2_{adj} = .29$. Both naturalness, $\beta = 0.50$, $t(625) = 13.55$, $p < .001$, $\eta_p^2 = .23$, and disorder, $\beta = -0.44$, $t(625) = -11.28$, $p < .001$, $\eta_p^2 = .17$, significantly predicted noun preference. A contrast indicated that the effect of naturalness on noun preference was significantly larger than the effect of perceived disorder, $F(1, 625) = 4.42$, $p = .036$, consistent with the nature-trumps-disorder hypothesis. We also calculated relative importance of naturalness and disorder for predicting noun preference as before. Across all eight metrics calculated, naturalness was estimated to be more important than disorder for noun preference—e.g., the recommended *lmg* method estimated that 58% of the variance in the model was explained by naturalness vs. 37% by disorder. Regarding the harmless-disorder and beneficial-disorder hypotheses, there was a significant negative interaction between the effects of naturalness and disorder on noun preference, $\beta = -0.18$, $t(625) = -5.06$, $p < .001$, $\eta_p^2 = .04$, contrary to the harmless-disorder and beneficial-disorder hypotheses. From these results, we conclude that semantics are also *sufficient* for the nature-trumps-disorder effect.

Chapter 2 Discussion

We drew several insights from these experiments. First, these experiments demonstrate that, when semantics are involved, nature trumps disorder in driving aesthetic preferences. This was evidenced across a series of experiments that utilized two completely different sets of

images of environmental scenes (N = 1,176) that widely varied in semantic content, and a set of nouns (N = 632) that also varied widely in semantic content. Interestingly, the nature-trumps-disorder effect was smaller in Experiment 4a-c than in Experiments 1a-c and 2a-c. This suggests that semantics alone may be sufficient to cause the nature-trumps-disorder effect, but the effect is strongly amplified by the visual features of nature scenes, consistent with past research that suggests that part of naturalness is determined by low-level visual features (e.g., Berman et al., 2014; Ruderman & Bialek, 1994; Torralba & Oliva, 2003), as well as part of nature's aesthetics (Kardan, Demiralp, et al., 2015). Second, there was no nature-trumps-disorder effect in Experiments 3a-f, in which we extracted and scrambled the basic visual features of the scenes. This suggests that basic visual features alone are *not sufficient* to cause the nature-trumps-disorder effect. This may be because naturalness has a weaker presence at the basic visual level compared to disorder. There is, however, nuance to this finding. Comparing results from Experiments 3a-c and 3d-f, it seems that naturalness may have been preserved more in colors than in edges, whereas disorder was preserved more in edges than in colors—an intriguing possibility that requires further research. Bringing in results from Experiments 4a-c, in which word stimuli yielded the effect, we concluded that visual features are also *not necessary*. Therefore, basic visual features are *neither necessary nor sufficient* for the nature-trumps-disorder effect, but they seem to amplify the effect at the level of scenes. In contrast, semantics are *both necessary and sufficient* for the nature-trumps-disorder effect.

More generally, this chapter presents and resolves the paradoxical relationship between naturalness and disorder. Much previous research has focused 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) but, to our knowledge, no research has systematically investigated the separate roles of naturalness and disorder on aesthetic preference. The finding that these environmental dimensions have independent effects suggests that there may be a fruitful avenue of research at the intersection of these two dimensions, which have been investigated in isolation. By showing that disorder matters in natural environments, we are extending disorder research into a whole new class of environments. The independence of the effects on aesthetic preference suggests that they may have other separable psychological effects. For example, a disordered natural environment may be restorative, but at the same time it may encourage rule-breaking (see Chapter 1).

This study has implications for other lines of research. If the high-level semantics of nature have strong affective importance tied to them, it may be more difficult to make 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). The importance of semantics in the tested natural scenes and the generally stronger effects of naturalness (e.g., compared to disorder in our study) could be related to its unique ties with dimensions with an evolutionary basis such as survivability (e.g., Nairne, Pandeirada, & Thompson, 2008; E. O. Wilson, 1984). This too is an area worthy of further inquiry.

This chapter also has practical significance. 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 linked to nature's restorative potential,

perhaps aesthetic preferences should be weighted even more. As the world becomes more populated and urbanized, there is a pressing demand to incorporate nature into built environments. Not only is it aesthetically pleasing, it is also economically sensible—according to a report by Booz Allen Hamilton (Booz Allen Hamilton, 2015), from 2015-2018, green construction is predicted to generate $303.4 billion in GDP, support 3.9 million jobs, and provide $268.4 billion in labor earnings. In addition, as virtual reality becomes more of a reality, there is a growing interest in designing salubrious virtual environments. This chapter directly suggests that order should be considered in the design of such greenspaces and virtual environments.

CHAPTER 3: THE PRESERVATION OF SEMANTICS IN VISUAL FEATURES

A scene of an environment contains a lot of information that we perceive as "features," broadly construed. There are lower-level visual features such as edges and colors and higher-level semantic features such as recognizable objects, places, and descriptors (Oliva & Torralba, 2001; Rosch & Mervis, 1975). Here we focus on two specific semantic features of a scene—its level of 'disorder' and its level of 'naturalness'—due to their psychological importance (Berman et al., 2008; Kotabe et al., 2016b; E. O. Wilson, 1984; J. Q. Wilson & Kelling, 1982). Traditional non-Gestalt visual perception models suggest that integration of low-level visual features and segmentation of the scene must occur *before* high-level semantic features are perceived (e.g., Biederman, 1987; Treisman & Gelade, 1980; David Marr, 1976). This would imply that low-level visual features do not intrinsically carry information about high-level semantic features. Here we question this assumption by asking, can the low-level visual features of a scene preserve any of the high-level semantics of that scene? Furthermore, is it possible that different high-level semantics are preserved in different types of low-level visual features?

The preservation of high-level semantics in low-level visual features would be of import to theories of visual perception, posing a challenge especially to those that assume that semantic processing starts later in visual perception. First, it would suggest that semantic processing may start earlier than thought previously. Second, it would suggest that integration of low-level visual features and scene segmentation may occur after semantic processing has begun, or in parallel.

We know of some of work that is relevant to this idea. First, although it may seem improbable that humans can start to process semantics from information carried by low-level visual features, before objects are perceived, we invoke the argument that the brain is a meaning

making machine that can even find meaningful objects in white noise (Gosselin & Schyns, 2003). Furthermore, there is ample evidence that people can rapidly identify the semantic category of a scene—a remarkable feat considering the subtle comparisons one must make among the large number of scenes within a scene category (e.g., imagine the number of scenes one could consider 'natural'), not to mention the large number of scene categories. This research suggests that object perception is not necessary to identify the semantic category of a scene (Oliva & Torralba, 2006). After only 20 ms of exposure to a scene, people can categorize whether the scene contains an animal or not with about 94% accuracy (Thorpe, Fize, & Marlot, 1996). This is a shorter duration than used in some subliminal priming experiments! After only 27 ms of exposure to a scene, people can recall seeing semantic features, as evidenced by a free-recall experiment (Fei-Fei, Iyer, Koch, & Perona, 2007). After only 33 ms of exposure to a scene, people can not only categorize objects in a scene (e.g., dog) but can even identify within-category kinds (e.g., a German Shepherd) above chance (Grill-Spector & Kanwisher, 2005). Even if scenes are jumbled into six parts and presented for only 50 ms, people can categorize the gist of the scenes better than chance (Biederman, Rabinowitz, Glass, & Stacy, 1974). After 100 ms of exposure to a scene, people can perceive if an object is incompatible within the scene (Biederman, Teitelbaum, & Mezzanotte, 1983).

There is other support for our hypothesis. An electroencephalogram (EEG) experiment showed that low-level category-dependent processing can occur within about 75 ms after the 20 ms presentation of a stimulus (Vanrullen & Thorpe, 2001). Not only was the presentation rapid, but category-dependent brain processing started soon after exposure, consistent with low-level visual information carrying semantic information. There are studies that suggest that people can

identify the semantic category of a scene in the near absence of attention (Fei-Fei, VanRullen, Koch, & Perona, 2005; Li, VanRullen, Koch, & Perona, 2002). At least one study suggests that it takes the same amount of time to detect an object as it does to categorize it (Grill-Spector & Kanwisher, 2005). This finding is inconsistent with the idea that semantic processing starts at a higher and more time-delayed level. An fMRI study also supports this idea by showing that scene categories could be decoded from activity in V1 (Walther et al., 2009). In fact, the decoding accuracy of V1 (26%) was not that far off from the decoding accuracy of the parahippocampal place area (31%), which is known to be a key region involved in processing scene semantics.

Specifically concerning the preservation of semantics in low-level spatial features, Oliva and Torralba (2001) presented the spatial envelope model which proposes that the global spatial layout of a scene, defined by specific low-level visual feature configurations, carries information about the semantic category (e.g., natural vs. built) of that scene (see also Oliva & Torralba, 2006). This computational model suggests that segmentation and the processing of individual objects or regions is not necessary for classifying scenes into semantic categories.

As for the preservation of semantics in low-level color features, although some research suggests that color information is not critical for the rapid categorization of scenes (Delorme, Richard, & Fabre-Thorpe, 2000; Fei-Fei et al., 2005), other research suggests otherwise. Oliva and Schyns (2000) showed that color information helps people categorize scenes into semantic categories when the color information is diagnostic of a semantic category. Follow-up research by Goffaux et al. (2005) provided both behavioral and EEG evidence that diagnostic color information is part of the scene "gist" (Oliva, 2005) that facilitates rapid scene recognition. This

is consistent with other research that suggests that prior experience benefits rapid scene understanding (Greene, Botros, Beck, & Fei-Fei, 2015). In fact, Goffaux et al. (2005) showed that atypical scene colors hinder rapid scene recognition.

Specifically concerning the preservation of semantics related with disorder and naturalness in low-level visual features, we showed that the disorder of a scene could be predicted by objective low-level visual features in Chapter 1 (see also Kotabe et al., 2016b), and we have shown that this is also true for naturalness (Berman et al., 2014). Relatedly, Oliva and Torralba (2001) showed that naturalness could be predicted based on the principal components of power spectra which capture orientation and spatial frequency information. It is unclear, however, whether this is possible because the disorder and naturalness of a scene systematically varies non-causally with certain low-level visual features (e.g., the low-level visual features relate with objects that convey semantics related with disorder or naturalness), or if the preservation of high-level semantics in low-level visual features plays a role. Here we test the latter possibility.

Notes on General Method

We sampled broadly from real-world environments by utilizing both the set of 260 scene images described in Chapter 1 and the set of 916 scene images described in Chapter 2. We manipulated these scene images by extracting and scrambling their low-level edge features and their low-level color features. We collected disorder and naturalness ratings for these stimuli. Data analysis was conducted on the image-level summary statistics.

Note that we did not use the rapid scene recognition paradigm for this study. Although this paradigm is useful for the study of how much time it takes to extract certain semantic

information from a scene, it does not directly test whether low-level visual information carries high-level semantic information. It also relies on recognition instead of directly testing perception. The method we used, which involved freely rating semantic dimensions of presented scenes, directly measured the perception of these semantic dimensions. Furthermore, by taking these measurements between-subjects, we eliminated memory issues including the possibility that high-level semantics are preserved in low-level visual features only when one has previously viewed the unaltered scene (thus has memory of the scene and low-level visual features), or when one has previously viewed its low-level visual features in a scrambled stimulus (thus has memory of the low-level visual features).

Experiment 1a-b: Is Disorder Preserved in Edges?

We extracted and scrambled the edges and colors of the 916 scene images described in Chapter 2 using edge and color extraction and scrambling methods as in Chapter 1. We had people rate the scrambled-edge stimuli (Experiment 1a) and the scrambled-color stimuli (Experiment 1b) in terms of disorder. We then tested the association of these disorder ratings with the disorder ratings of the original scenes (previously collected ratings, see Chapter 2) to see if disorder was preserved in the low-level visual edge or color features. Note that this experiment was a replication of Experiment 2 in Chapter 1 with a larger and more diverse set of scene images, but we conducted additional data analysis and further interpreted the results in this context. Based on the results of Experiment 2 in Chapter 1, we predicted that the disorder ratings of the scrambled-edge stimuli would correlate stronger with the disorder ratings of the original scenes than would the disorder ratings of the scrambled-color stimuli.

Method

Participants and design. 221 US-based adults (122 men, 98 women, 1 other) were recruited from AMT and participated in Experiment 1a (scrambled-edges). 241 US-based adults (128 men, 112 women, 1 other) were recruited from AMT and participated in this Experiment 1b (scrambled-colors). Sample size and stopping rule were based on our goal to receive ~20 ratings per image. Across experiments, ages ranged from 19 to 73 (M = 34.77, SD = 10.76). 356 participants identified primarily as White/Caucasian, 34 as Asian/Asian American, 33 as Black/African American, 24 as Hispanic/Latino, 9 as multiple ethnicities, 3 as Native American/Alaska Native, and 3 as Native Hawaiian/Pacific Islander. The median experiment duration was 8 minutes and participants were compensated $1.50 for participating. Informed consent was administered by the IRB of the University of Chicago.

Procedure. Participants were first given a brief introduction to the image-rating task. They were instructed, "You will be presented with a series of 100 images containing various lines (colors). We simply want you to rate each image in terms of how disorderly or orderly it looks." Participants were then randomly presented 100 of the 916 scrambled-edge stimuli (Experiment 1a) or scrambled-color stimuli (Experiment 1b) on a plain white background (they rated scrambled-edge and scrambled-color stimuli derived from the full set of 1,105 images but 189 of these images and their corresponding visual derivations were removed from analysis as in Chapter 2 because they contained social disorder as assessed by the rule-breaking ratings). The randomization scheme had two layers. First, we randomly selected 20 images from each quintile of urbanness/naturalness. Second, we presented these 100 images in random order. This ensured that each participant would view a wide sample of images from more urban to more natural. For

each image, they were instructed to rate the scene in terms of disorder on a seven-point semantic differential scale ranging from “very disorderly” to “very orderly.” The task would continue to the next image immediately after a rating was made. By not fixing presentation time, we would not artificially make people view the scenes for shorter or longer than they wanted to, which could have influenced their perceptions.

Results

We correlated the disorder ratings of the scrambled-edge and scrambled-color stimuli with the previously collected disorder ratings of the original scenes. Disorder ratings of the scrambled-edge stimuli significantly correlated with disorder ratings of the original scenes, $r = .31$, $p < .001$, providing first evidence that disorder was partially preserved in the low-level edge features (see Figure 11a). In contrast, disorder ratings of the scrambled-color stimuli significantly correlated *negatively* with disorder ratings of the original scenes, $r = -.15$, $p < .001$ (see Figure 11b), suggesting that scene-level disorder was not preserved in the color features. The difference between these two dependent correlations was statistically significant, $t = 9.74$, $p < .001$, according to Williams’ test (1959b).

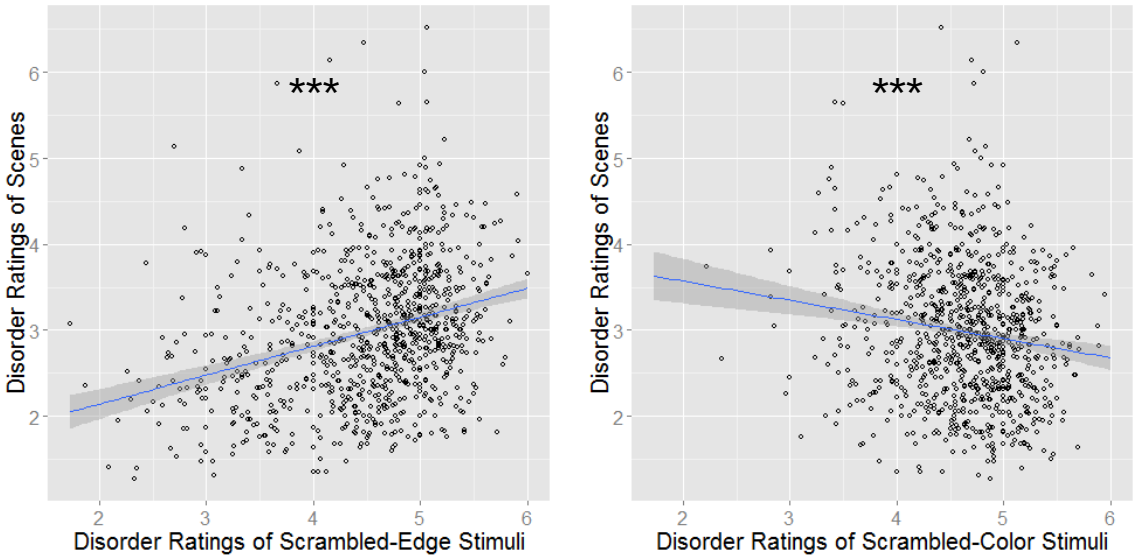


Figure 11: Results of Experiment 1. (a) Disorder ratings of scrambled-edge stimuli significantly correlated with the disorder ratings of scene images. (b) Disorder ratings of scrambled-color stimuli significantly correlated *negatively* with the disorder ratings of scene images. Least-squares lines with 95% confidence bands shown. *** $p < .001$.

Because of imperfect linearity, we also tested these associations with two nonparametric tests of association based on rank-order, Spearman's rho (ρ) and Kendall's tau-b (τ). Disorder ratings of the scrambled-edge stimuli were again significantly associated with disorder ratings of the original scenes according to both tests, $\rho = .32$, $p < .001$ and $\tau = .22$, $p < .001$, providing further evidence that disorder was partially preserved in the low-level edge features. In contrast, disorder ratings of the scrambled-color stimuli were again significantly correlated *negatively* with disorder ratings of the original scenes according to both tests, $\rho = -.12$, $p < .001$ and $\tau = -.08$, $p < .001$, again suggesting that scene-level disorder was not preserved in the color features. The difference between the two dependent ρs was statistically significant, $t = 10.22$, $p < .001$, according to Williams' test.

These results suggest that high-level semantics related to disorder at the scene-level were preserved in the low-level edge features of the scenes but not as much in the low-level color features of the scenes. This is direct evidence that high-level semantics can be preserved in low-level visual features, and more specifically, that some types of low-level visual features carry certain semantic information better than others. But is it possible that different semantic information is preserved better in different low-level visual features? Specifically, is 'naturalness' better preserved in colors than in edges because colors are more diagnostic of naturalness (Oliva & Schyns, 2000)? We tested this possibility in the following experiment. We note that this would be contrary to the spatial envelope model (Oliva & Torralba, 2001) which suggests that naturalness is a perceptual dimension that is well-represented by the *spatial* structure of a scene.

Experiment 2: Is Naturalness Preserved in Colors?

We extracted and scrambled the edges and colors of the 260 scene images described in Chapter 1 using color extraction and scrambling methods as in Chapter 1. We had people rate these derived stimuli in terms of naturalness. We then tested the association of these naturalness ratings with the naturalness ratings of the original scenes to see if naturalness was preserved in the low-level visual edge or color features. We predicted that the naturalness ratings of the scrambled-color stimuli would correlate stronger with the naturalness ratings of the original scenes than would the naturalness ratings of the scrambled-edge stimuli, under the assumption that colors are more diagnostic of naturalness.

Method

Participants and design. 186 US-based adults (118 men, 67 women, 1 other) were recruited from AMT and participated in this two-condition (stimuli: scrambled edges vs. scrambled colors) between-subjects experiment. Sample size and stopping rule were based on our goal to receive ~20 ratings per image. Ages ranged from 18 to 62 (M = 32.03, SD = 9.59). 138 participants identified primarily as White/Caucasian, 20 as Asian/Asian American, 14 as Black/African American, 7 as Hispanic/Latino, 4 as other, 1 as Native American/Alaska Native, and 1 as Native Hawaiian/Pacific Islander. The median experiment duration was 4 minutes and 9 seconds and participants were compensated $0.50 for participating. Informed consent was administered by the IRB of the University of Chicago

Procedure. The procedure was the same as in Experiment 1 except that participants rated naturalness on a seven-point semantic differential scale ranging from "very urban" to "very natural."

Results

The analysis followed the same procedure as in Experiment 1. Naturalness ratings of the scrambled-color stimuli significantly correlated with naturalness ratings of the original scenes, r = .24, p < .001, providing first evidence that naturalness was partially preserved in the low-level color features (see Figure 12c). In contrast, naturalness ratings of the scrambled-edge stimuli *did not* significantly correlate with naturalness ratings of the original scenes, r = -.06, p = .358 (see Figure 12a), suggesting that naturalness was not preserved as much in the edge features. In support, the difference between these two dependent correlations was statistically significant, t = 3.52, p < .001, according to Williams' test.

There was a clustering of scenes rated as highly natural (see Figure 12). After removing these with a cutoff of 6.5/7.0 on the naturalness scale (N = 212 remaining), the results provide even stronger support for our hypothesis. Naturalness ratings of the scrambled-color stimuli significantly correlated with naturalness ratings of the original scenes, $r = .44$, $p < .001$ (see Figure 12d). In contrast, naturalness ratings of the scrambled-edge stimuli *did not* significantly correlate with naturalness ratings of the original scenes, $r = .02$, $p = .358$ (see Figure 12b), suggesting that naturalness was not preserved as much in the edge features. In support, the difference between these two dependent correlations was statistically significant, $t = 4.87$, $p < .001$, according to Williams' test.

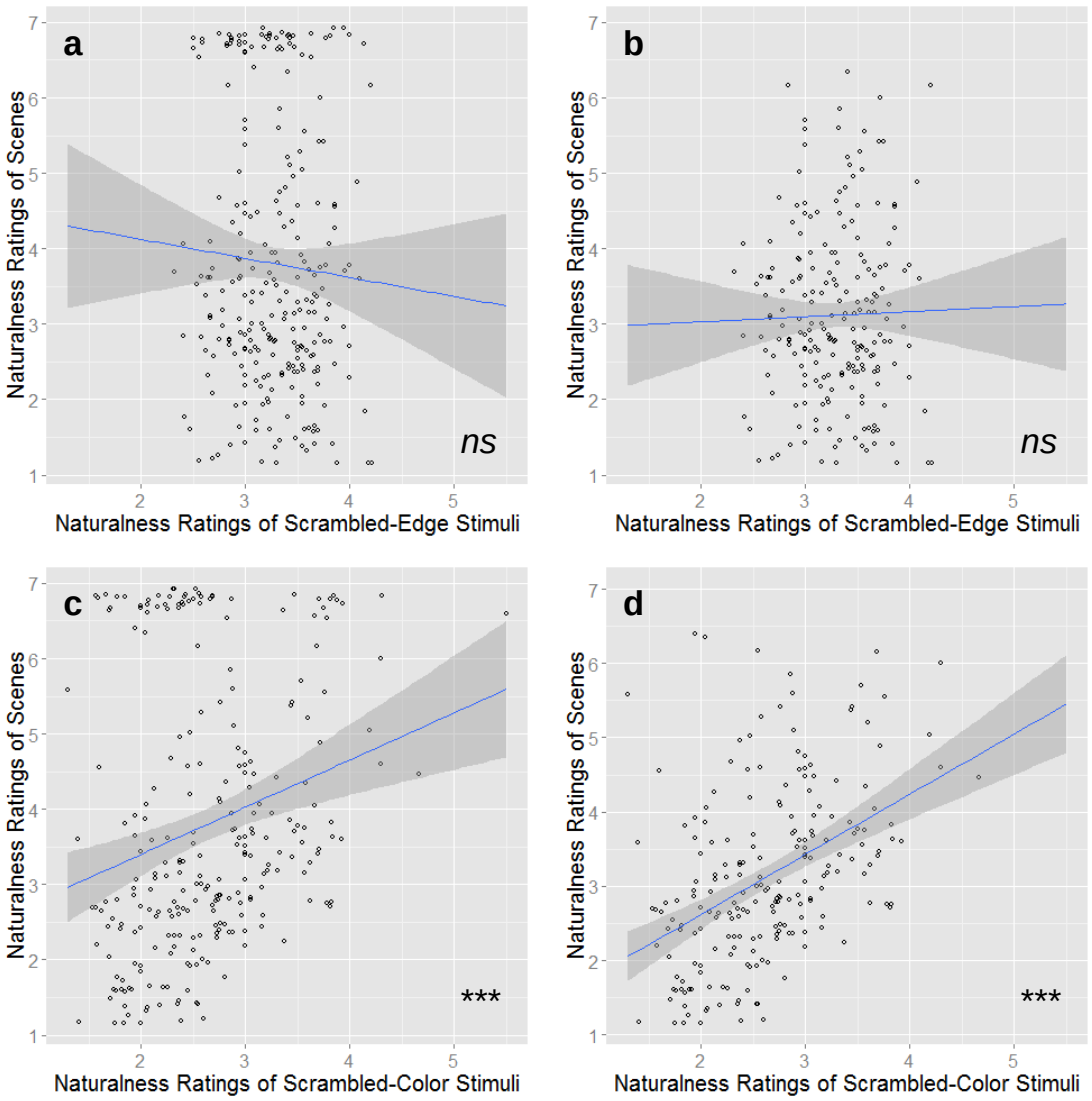


Figure 12: Results of Experiment 2 before and after removing cluster of highly natural scenes. (a-b) Naturalness ratings of scrambled-edge stimuli did not significantly correlate with the naturalness ratings of scene images; (c-d) Naturalness ratings of scrambled-color stimuli significantly correlated with the naturalness ratings of scene images. Least-squares lines with 95% confidence bands shown. *** $p < .001$, *ns* = not significant

Because of imperfect linearity, we also tested these associations with Spearman's rho and Kendall's tau-b. Naturalness ratings of the scrambled-color stimuli were again significantly associated with naturalness ratings of the original scenes according to both tests, before, $\rho = .29$,

$p < .001$ and $\tau = .21$, $p < .001$, and after, $\rho = .48$, $p < .001$ and $\tau = .34$, $p < .001$, removing the cluster of highly natural scenes, providing further evidence that naturalness was partially preserved in the low-level color features. In contrast, naturalness ratings of the scrambled-edge stimuli were again *not* significantly associated with naturalness ratings of the original scenes according to both tests, before, $\rho = -.07$, $p = .288$ and $\tau = -.05$, $p = .287$, and after, $\rho = -.01$, $p = .851$ and $\tau = -.01$, $p = .879$, removing the cluster of highly natural scenes, again suggesting that naturalness was not preserved as much in the edge features. In support, the difference between the two dependent ρs was statistically significant before, $t = 4.27$, $p < .001$, and after, $t = 5.74$, $p < .001$, removing the highly natural scenes, according to Williams' test.

These results suggest that high-level semantics related to naturalness at the scene-level were preserved in the color features of the scenes but not as much in the edge features of a scene. This further supports our general hypothesis that high-level semantics can be preserved in low-level visual features. It also further supports our more specific hypothesis that some low-level visual features carry certain semantic information better than others.

Chapter 3 Discussion

Together, the experiments of Chapter 3 provide direct evidence that high-level semantics can be preserved in low-level visual features, and that different high-level semantics can be preserved in different types of low-level visual features. This is evidenced by our two experiments, the first showing that high-level semantics related with disorder were preserved better in low-level edge features than in low-level color features, and the second showing that high-level semantics related with naturalness were preserved better in low-level color features than in low-level edge features. This research adds to the body of literature that is starting to

entertain the possibility that object perception and segmentation do *not* need to occur before identifying the semantic category of a scene.

SUMMARY AND CONCLUSIONS

In Chapter 1, we focused on disorderly environments and their influence on rule-breaking behaviors. Specifically, we were concerned with the confounding of visual disorder and social disorder in previous research bearing on broken windows theory. This led us to wonder whether visual disorder alone, in the absence of social disorder cues, could encourage rule-breaking behavior. To that end, we conducted several experiments to deconstruct and define visual disorder. Our results indicated the visual disorder is more a function of spatial features (particularly density of curved edges and asymmetry) than color features. In a following set of experiments, we reconstructed visual disorder based on what we learned in the previous experiments, then tested whether visually disordered stimuli alone could encourage a typical rule-breaking behavior—cheating. We found that subtly manipulating visual disorder increased the likelihood of cheating by up to 35% and the average magnitude of cheating by up to 87%. In a final experiment (Experiment 7), we explored potential mechanisms of these effects by analyzing what people wrote about after viewing the visually-ordered vs visually-disordered stimuli. The results from this exploratory experiment provided initial evidence that there could be some information "overload" mechanism or some priming/spreading-activation mechanism at work—or perhaps both. A key takeaway from this work is that explanations for traditional theories of rule-breaking that assume social disorder cues and complex social reasoning are necessarily involved, should be reconsidered. These experiments also add to a growing literature that suggests that simple perceptual properties of the environment are involved in high-level psychological processes that may have downstream effects on complex behaviors. They also bear on the question of to what extent our actions are within our control.

In Chapter 2, we focused on disorder, naturalness, and their joint influence on aesthetic preferences. Specifically, we were interested in why nature scenes are aesthetically preferred when they are also disorderly, when various research suggests that disorder is aesthetically aversive. We tested three competing hypotheses. First, it is possible that the effect of naturalness on aesthetic preference trumps the effect of disorder on aesthetic aversion, and that these effects are independent of each other (*nature-trumps-disorder hypothesis*). Second, it is possible that disorder simply does not matter in natural contexts (*harmless-disorder hypothesis*). Third, it is possible that disorder is actually aesthetically preferable in natural contexts (*beneficial-disorder hypothesis*). Our results lend support to the nature-trumps-disorder hypothesis and are contrary to the other hypotheses. Further, they suggest that nature semantics are particularly important for the nature-trumps-disorder effect. Specifically, nature semantics were shown to be both necessary and sufficient for the nature-trumps-disorder effect, and their interaction with low-level visual features amplified the effect. At a broader level, this means that recognizable entities in a scene can suppress or strengthen the relationships between naturalness, disorder, and aesthetic preferences. Further, these results bear on psychological theories concerning the joint influence of lower-level visual features, higher-level semantics, and their interactive effects on affect and cognition.

In Chapter 3, we focused on whether disorder and naturalness semantics of a scene could be preserved in the low-level visual features of that scene. If so, this would stand in contrast to traditional visual perception models that suggest that integration of low-level visual features and segmentation of the scene must occur before high-level semantic features are perceived. This traditional view implies that low-level visual features of a scene alone would not carry semantic

information related to that scene. Our results not only indicate that low-level visual features of a scene alone can carry semantic information related to that scene, but also that different high-level semantics can be preserved in different types of low-level visual features. Specifically, the disorder semantics of a scene are better preserved in edge features than color features, whereas the naturalness semantics of a scene are better preserved in color features than edge features. These findings suggest that semantic processing may start earlier than thought before, and integration of low-level visual features and segmentation of the scene may occur after semantic processing has begun, or in parallel.

These chapters present a variety of evidence substantiating the intimate connection we have to our ever-present physical environments, and provide insight into the linkage between low-level visual processing, semantics, aesthetic sense, and behavior. They not only have implications for psychological theories on these phenomena, and their interactions, but also implications for environmental design. The research in Chapter 1 provides a quantification of elements of visual disorder and shows that visual disorder can have marked effects on rule-breaking behavior. It may be fruitful if environmental designers start to weight visual order more in their design decisions, and our research provides some tangible ways to think about manipulating visual disorder. Of course, there are many remaining questions about the components of visual disorder, and the components of environmental disorder more broadly, and answering these questions will further benefit the informed design of environments. The research in Chapter 2 is particularly relevant to the design of greenspaces. It suggests that more orderly green space is aesthetically preferable to more disorderly green space, contrary to evolutionary theorizing that suggests the opposite. However, there is bound to be nuance to this topic. A quick

Google Image search of "beautiful architecture" reveals buildings that have elements of order such as symmetry but also elements of a sort of naturalistic disorder. The research in Chapter 3 also has design relevance. It suggests that edges and colors alone can convey semantic information and thus semantics should be considered even in design decisions about such low-level visual features. This is a huge area of research that we only touched on by focusing on the preservation of disorder and naturalness in edges and colors. And although our results indicate that global scene-semantic information can be carried by low-level visual features, they do not tell us about whether more local object-semantic information can be carried by such features. As a whole, this dissertation builds on a science that finds the human eternally bonded to their physical environment, reciprocally—and predictably—shaping each other.

REFERENCES

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235.

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.

Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: "Liking","wanting", and learning. *Current Opinion in Pharmacology*, *9*, 65–73.

Bertrand, F., Meyer, N., & Maumy-Bertrand, M. (2014). *Partial Least Squares Regression for Generalized Linear Models, R package version 1.1.1*.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, *103*(3), 597.

Biederman, I., Teitelbaum, R. C., & Mezzanotte, R. J. (1983). Scene perception: A failure to find a benefit from prior expectancy or familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(3), 411.

Booz Allen Hamilton. (2015). *U.S. Green Building Council: Green building economic impact study*.

Braga, A. A., & Bond, B. J. (2008). Policing crime and disorder hot spots: A randomized controlled trial. *Criminology*, *46*, 577–607.

Braga, A. A., Weisburd, D. L., Waring, E. J., Mazerolle, L. G., Spelman, W., & Gajewski, F. (1999). Problem-oriented policing in violent crime places: A randomized controlled experiment. *Criminology*, *37*, 541–580.

Brascamp, J., Blake, R., & Knapen, T. (2015). Negligible fronto-parietal BOLD activity accompanying unreportable switches in bistable perception. *Nature Neuroscience, 18*(11), 1672–1678.

Bratman, G. N., Daily, G. C., Levy, B. J., & Gross, J. J. (2015). The benefits of nature experience: Improved affect and cognition. *Landscape and Urban Planning*, *138*, 41–50.

Bratman, G. N., Hamilton, J. P., Hahn, K. S., Daily, G. C., & Gross, J. J. (2015). Nature experience reduces rumination and subgenual prefrontal cortex activation. *Proceedings of the National Academy of Sciences*, *112*(28), 8567–8572.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.

Casasanto, D., & Bottini, R. (2014). Spatial language and abstract concepts. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(2), 139–149.

Chae, B. G., & Zhu, R. J. (2014). Environmental disorder leads to self-regulatory failure. *Journal of Consumer Research*, *40*, 1203–1218.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*, 497–505.

Cutrona, C. E., Russell, D. W., Hessling, R. M., Brown, P. A., & Murry, V. (2000). Direct and moderating effects of community context on the psychological well-being of African American women. *Journal of Personality and Social Psychology*, *79*(6), 1088–1101.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: A study in monkeys and humans. *Vision Research*, *40*(16), 2187–2200.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*(2), 301–318.

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), 10.

Fei-Fei, L., VanRullen, R., Koch, C., & Perona, P. (2005). Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition*, *12*(6), 893–924.

Felleman, D. J., & Essen, D. C. V. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex*, *1*(1), 1–47.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, *4*(12), 2379–2394.

Fishbach, A., Friedman, R. S., & Kruglanski, A. W. (2003). Leading us not unto temptation: Momentary allurements elicit overriding goal activation. *Journal of Personality and Social Psychology, 84*, 296–309.

Geis, K. J., & Ross, C. E. (1998). A new look at urban alienation: The effect of neighborhood disorder on perceived powerlessness. *Social Psychology Quarterly, 61*, 232–246.

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

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363.

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P., & Rossion, B. (2005). Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Visual Cognition*, *12*(6), 878–892.

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science*, *14*(5), 505–509.

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2015). What you see is what you expect: Rapid scene understanding benefits from prior experience. *Attention, Perception, & Psychophysics*, *77*(4), 1239–1251.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, *16*(2), 152–160.

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.

Harcourt, B. E. (2009). *Illusion of order: The false promise of broken windows policing*. Harvard University Press.

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

Hauser, D. J., & Schwarz, N. (2015). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 1–8.

Heintzelman, S. J., Trent, J., & King, L. A. (2013). Encounters with objective coherence and the experience of meaning in life. *Psychological Science*, *24*, 991–998.

Hofmann, W., Friese, M., Schmeichel, B. J., & Baddeley, A. D. (2011). Working memory and self-regulation. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of self-regulation:*

Research, theory, and applications (Vol. 2, pp. 204–225). New York, NY: Guilford Press.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*, 399–425.

Humpel, N., Owen, N., & Leslie, E. (2002). Environmental factors associated with adults' participation in physical activity: A review. *American Journal of Preventive Medicine*, *22*, 188–199.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? (pp. 145–152). Presented at the Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE.

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.

Kardan, O., Gozdyra, P., Misic, B., Moola, F., Palmer, L. J., Paus, T., & Berman, M. G. (2015). Neighborhood greenspace and health in a large urban center. *Scientific Reports*, *5*, 11610.

Kardan, O., Henderson, J., Yourganov, G., & Berman, M. G. (in press). Observer's cognitive states modulate how visual inputs relate to gaze control. *Journal of Experimental Psychology: Human Perception and Performance*.

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, *322*, 1681–1685.

Kelling, G. L., & Coles, C. M. (1997). *Fixing broken windows: Restoring order and reducing crime in our communities*. New York, NY: Touchstone.

Kinchla, R. A. (1977). The role of structural redundancy in the perception of visual targets. *Perception & Psychophysics*, *22*(1), 19–30.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., … Brumbaugh, C. C. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152.

Kotabe, H. P. (2014). The world is random: A cognitive perspective on perceived disorder. *Frontiers in Psychology*, *5*.

Kotabe, H. P., & Hofmann, W. (in press). How depletion operates in an integrative theory of self-control. In *Self-Regulation and Ego Control*.

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? Manuscript submitted for publication.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2016b). *The order of disorder: Deconstructing visual disorder and its effect on rule-breaking*. Manuscript submitted for publication.

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

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, *99*(14), 9596–9601.

Linares, L. O., Heeren, T., Bronfman, E., Zuckerman, B., Augustyn, M., & Tronick, E. (2001). A mediational model for the impact of exposure to community violence on early child behavior problems. *Child Development, 72*, 639–652.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Glenview, IL: Scott, Foresman.

Mahoney, J. V. (1987). *Image chunking: Defining spatial building blocks for scene analysis*. DTIC Document.

Marr, D. (1976). Early processing of visual information. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *275*(942), 483–519.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London B: Biological Sciences*, *207*(1167), 187–217.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45*, 633–644.

McIntosh, A. R., Grady, C. L., Ungerleider, L. G., Haxby, J. V., Rapoport, S. I., & Horwitz, B. (1994). Network analysis of cortical visual pathways mapped with PET. *Journal of Neuroscience, 14*(2), 655.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*, 3–19.

Nairne, J. S., Pandeirada, J. N., & Thompson, S. R. (2008). Adaptive memory: The comparative value of survival processing. *Psychological Science*, *19*, 176–180.

Oliva, A. (2005). Gist of the scene. *Neurobiology of Attention*, *696*(64), 251–258.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, *41*(2), 176–210.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*, 1023–1031.

Perkins, D. D., & Taylor, R. B. (1996). Ecological assessments of community disorder: Their relationship to fear of crime and theoretical implications. *American Journal of Community Psychology*, *1*, 63–107.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*, 130–137.

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

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179.

Ransdell, S., Levy, C. M., & Kellogg, R. T. (2002). The structure of writing processes as revealed by secondary task demands. *L1-Educational Studies in Language and Literature*, *2*(2), 141–163.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Ross, C. E. (2000). Neighborhood disadvantage and adult depression. *Journal of Health and Social Behavior*, *41*, 177–187.

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters*, *73*, 814–817.

Sampson, R. J., & Raudenbush, S. W. (2004). Seeing disorder: Neighborhood stigma and the social construction of "broken windows." *Social Psychology Quarterly*, *67*, 319–342.

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.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York, NY: Teachers College, Columbia University.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

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

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Tullett, A. M., Kay, A. C., & Inzlicht, M. (2015). Randomness increases self-reported anxiety and neurophysiological correlates of performance monitoring. *Social Cognitive and Affective Neuroscience*, *10*, 628–635.

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

Vanrullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461.

Vohs, K. D., Baumeister, R. F., Mead, N. L., Hofmann, W., Ramanathan, S., & Schmeichel, B. J. (2013). Engaging in self-control heightens urges and feelings. \iManuscript submitted for publication.

Vohs, K. D., Redden, J. P., & Rahinel, R. (2013). Physical order produces healthy choices, generosity, and conventionality, whereas disorder produces creativity. *Psychological Science*, *24*, 1860–1867.

Wagner, D. D., Altman, M., Boswell, R. G., Kelley, W. M., & Heatherton, T. F. (2013). Self-regulatory depletion enhances neural responses to rewards and impairs top-down control. *Psychological Science*, *24*, 2262–2271.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *The Journal of Neuroscience*, *29*(34), 10573–10581.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect - the Panas Scales. *Journal of Personality and Social Psychology*, *54*, 1063.

Williams, E. J. (1959a). The comparison of regression variables. *Journal of the Royal Statistical Society: Series B*, *21*, 396–399.

Williams, E. J. (1959b). The comparison of regression variables. *Journal of the Royal Statistical Society: Series B*, *21*, 396–399.

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

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

Witkin, A. P., & Tenenbaum, J. M. (1983). On the role of structure in vision. In *Human and Machine Vision* (pp. 481–543).

Xiao, J., Hays, J., Ehinger, K., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo (pp. 3485–3492). Presented at the 2010 IEEE conference on Computer vision and pattern recognition (CVPR), IEEE.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*, 151–175.

APPENDIX: SUPPLEMENTARY MATERIALS

Contents:

1. Notes on Amazon Mechanical Turk (AMT) samples

Across the reported experiments, only participants with at least a 95% overall approval rating were allowed to participate. Overall approval rating is equal to the proportion of work done by AMT workers approved by AMT requestors. Research suggests that workers with at least a 95% approval ratings are more attentive than workers with lower approval ratings (Peer, Vosgerau, & Acquisti, 2014).

Although there is a cost of slightly less environmental control than a lab study, running studies on AMT has a number of advantages. For example, it provides easy, quick, and cheap access to large, diverse, and reliable samples (Horton et al., 2011; Rand, 2012). AMT samples are significantly more representative of the U.S. population than college student samples (Buhrmester, Kwang, & Gosling, 2011). Also, there is no person-to-person direct contact like in a laboratory experiment, which can introduce complications such as experimenter effects and interpersonal dynamics. Also, AMT participants may be more attentive than undergraduate participants in psychology studies (Hauser & Schwarz, 2015; Klein et al., 2014). Klein et al. showed that tests attentiveness to instructions at a significantly higher rate than undergraduates at 15 out of 19 college sites and descriptively higher than undergraduates at the four other college sites. Hauser and Schwarz (2015) also showed that AMT participants were significantly more attentive to instructions than college students, but extended this finding to novel instructional manipulation checks.

2. Chapter 1: Experiment 1 supplement

Figure S1: Disorder-urbanness matrix showing examples of the various 260 environmental images used in Experiment 1. Each row and column represents a quintile on the given variable. For the following experiments, we extracted and scrambled the edge and color features from the images to remove possible social cues.

Table S1 presents the correlations between the low-level visual features and disorder ratings.

	1	2	3	4	5	6	7	8	9	10
Disorder ratings (1)	–									
Non-straight edge density (2)	.29***	–								
Straight edge density (3)	-.15*	-.55***	–							
Asymmetry (4)	-.11	-.75***	.26***	–						
Hue (5)	.04	.43***	-.16**	-.41***	–					
Saturation (6)	-.09	.39***	-.04	-.36***	.21**	–				
Value (7)	.04	-.23***	.07	.10	-.15*	-.28***	–			
SD hue (8)	.08	-.40***	.06	.35***	-.03	-.53***	.38***	–		
SD saturation (9)	.00	.21**	.06	-.19**	.33***	.55***	-.22***	-.18**	–	
SD value (10)	.06	-.15*	-.15*	.30***	-.07	-.08	.08	.17**	.15*	–

Table S1: Experiment 1 correlation matrix. N = 260; *** p < .001, ** p < .01, * p < .05

Separately regressing disorder ratings on the spatial features and the color features to compare R^2_{adj}s. Non-straight edge density, $\beta = 0.53$, $t(256) = 4.98$, $p < .001$, $\eta_p^2 = .088$, and asymmetry, $\beta = 0.27$, $t(256) = 2.95$, $p = .003$, $\eta_p^2 = .033$, significantly predicted disorder ratings. Straight-edge density did not significantly predict disorder ratings. None of the color features significantly predicted disorder ratings.

3. Chapter 1: Experiment 2 supplement

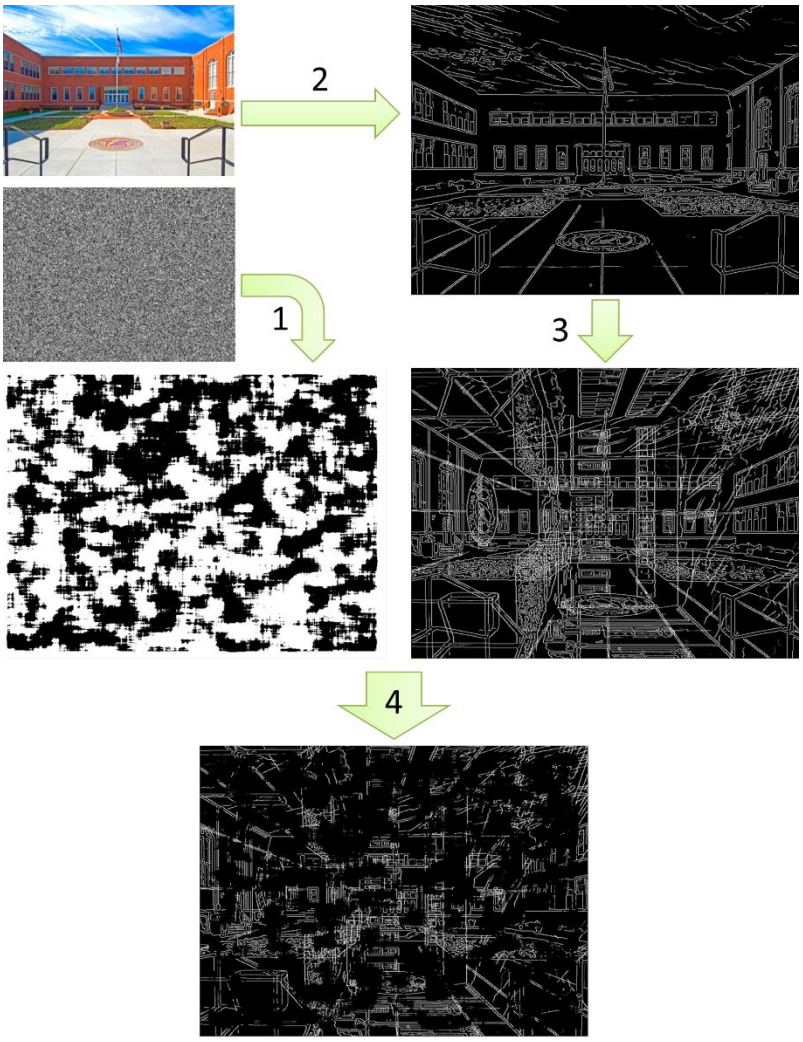


Figure S2: Illustration of the edge extraction and scrambling process. Process 1: Started with a mask matrix constructed to be the same size as the scene images (600*800) with its elements randomly assigned between zero and one. This matrix was then convolved with a median filter sized 30*40 pixels. 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 half of every mask having, on average, half a surface of 1s and half a surface of 0s. Process 2: Created edge map from the original image. Process 3: The edge map was randomly rotated either 90 or 270 degrees and overlaid on the 180-degrees-rotated edge map creating a stimulus comprising twice as many edges (but same straight and non-straight edge ratios) as the scene image. Process

4: This stimulus was then multiplied (dot product) by the mask so that half of its edges got removed at random.

Table S2 presents the correlations between the disorder ratings for the stimuli used in Experiment 2.

	DR-original	DR-edges	DR-colors
Disorder ratings for original environmental images (DR-original)	–		
Disorder ratings for scrambled-edge stimuli (DR-edges)	.38***	–	
Disorder ratings for scrambled-color stimuli (DR-colors)	.02	.02	–

Table S2: Experiment 2 correlation matrix. N = 260; *** p < .001.

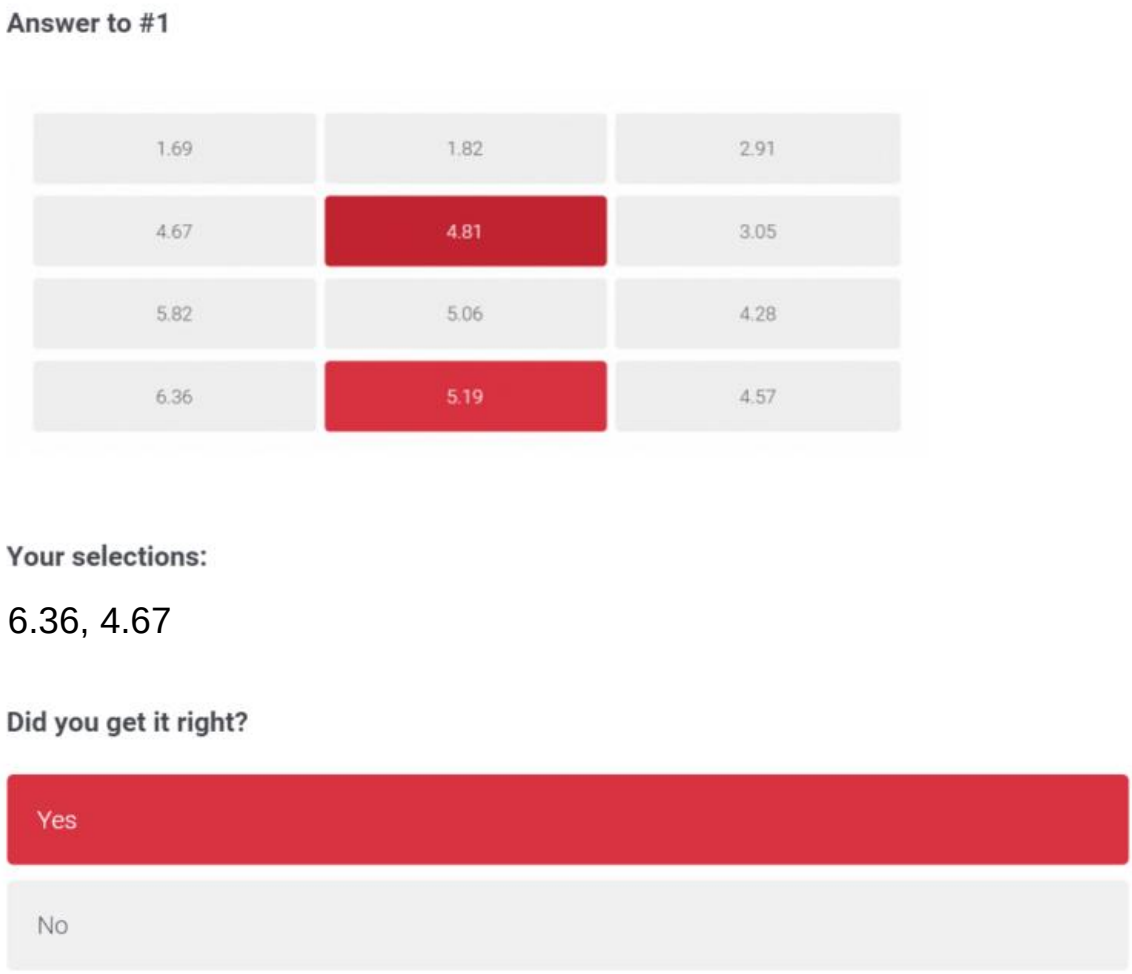
4. Chapter 1: Experiment 3 supplement

Table S3 presents the correlations between the disorder ratings for the stimuli used in Experiment 3.

	DR-original	DR-CC	DR-CI	DR-C
Disorder ratings for original environmental images (DR-original)	–			
Disorder ratings for color-congruent stimuli (DR -CC)	.20**	–		
Disorder ratings for color-incongruent stimuli (DR -CI)	.18**	.42***	–	
Disorder ratings for control stimuli (DR -C)	.16**	46***	.43***	–

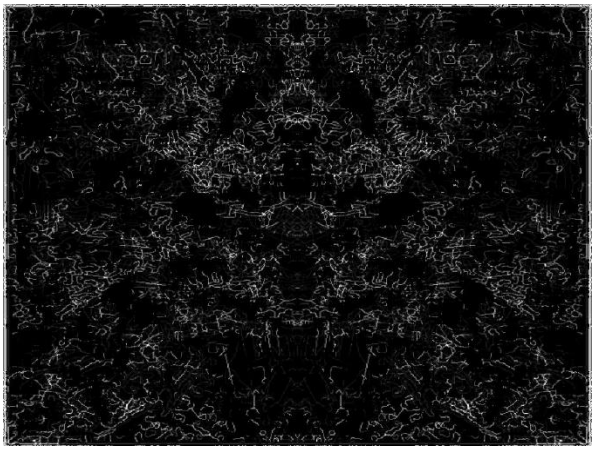
Table S3: Experiment 3 correlation matrix. N = 260; *** p < .001, ** p < .01

5. Chapter 1: Screenshot showing how people cheated in Experiments 4-6

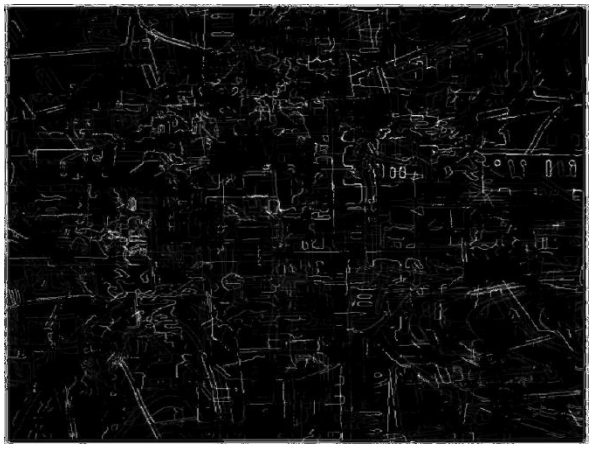


6. Chapter 1: Experiments 4-6 supplement

Symmetry + Visually Ordered Edges



Symmetry + Visually Disordered Edges



Asymmetry + Visually Ordered Edges

Asymmetry + Visually Disordered Edges

Figure S3: Examples of 2 (symmetry vs. asymmetry) × 2 (visually ordered edges vs. visually disordered edges) stimuli pretested in Experiment 4. We created and pretested 200 of such stimuli (50 from each cell).

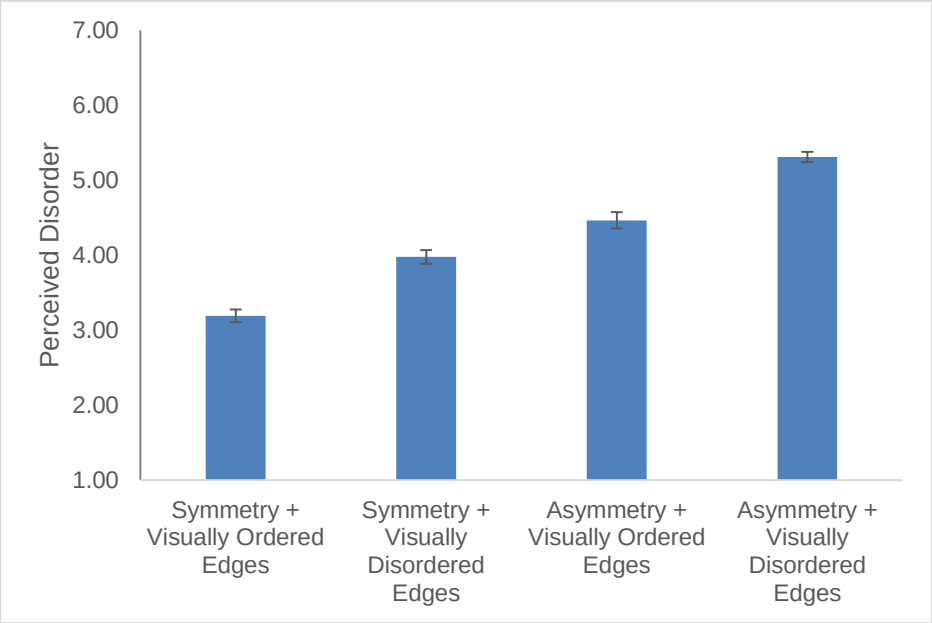


Figure S4: Mean disorder ratings for stimuli pretested for Experiment 4. The 30 most disorderly (all from asymmetry + visually disordered edges) and orderly (all from symmetry + visually ordered edges) stimuli were used for our manipulation of visual disorder vs. visual order in Experiments 4-7. Error bars indicate 95% CI of a one-sample t-test.

7. Chapter 1: Experiment 7 supplement

Here we present the stop words removed in the text analysis. Non-bold stop words are included in the *tm* R package and bold stop words were customly added.

i	me	my	myself	We	our	ours	ourselves	you	your
yours	yourself	yourselves	he	Him	his	himself	she	her	hers
herself	it	its	itself	They	them	their	theirs	themselves	what
which	who	whom	this	That	these	those	am	is	are
was	were	be	been	being	have	has	had	having	do
does	did	doing	would	should	could	ought	i'm	you're	he's
she's	it's	we're	they're	i've	you've	we've	they've	i'd	you'd
he'd	she'd	we'd	they'd	i'll	you'll	he'll	she'll	we'll	they'll
isn't	aren't	wasn't	weren't	hasn't	haven't	hadn't	doesn't	don't	didn't
won't	wouldn't	shan't	shouldn't	can't	cannot	couldn't	mustn't	let's	that's
who's	what's	here's	there's	when's	where's	why's	how's	a	an
the	and	but	if	Or	because	as	until	while	of
at	by	for	with	about	against	between	into	through	during
before	after	above	below	To	from	up	down	in	out
on	off	over	under	Again	further	then	once	here	there
when	where	why	how	All	any	both	each	few	more
most	other	some	such	No	nor	not	only	own	same
so	than	too	very	**also**	**sort of**	**just**	**kind of**	**seem**	**look**
like	**imag**	**pictur**	**photo**	**didnt**	**realli**				

8. Chapter 2: Experiments 4a-c supplement

abdomen	abode	academy	acid	acme
acorn	acre	aircraft	airport	airship
aisle	alarm	alcove	alley	almond
altar	altitude	amber	anchor	angel
anger	angle	animal	ankle	aperture
apple	apricot	apron	aquarium	arch
area	arena	aroma	arrow	asphalt
assembly	asylum	atom	attic	autumn
avenue	bacteria	badger	bakery	balcony
ballad	bamboo	bank	banquet	barn
barrel	basement	bathroom	battery	beach
beacon	bean	bear	beast	beaver
bedding	bedroom	beer	berry	bird
blizzard	blossom	blue	boar	boat
bonfire	bookcase	booth	border	bread
breeze	brewery	brick	bridge	broccoli
bronze	brown	bubble	buffalo	buffet
bull	bump	bundle	bunny	bush
business	butter	cabbage	cabin	cabinet
cable	calf	camel	camp	campus
canal	cane	canoe	canopy	canteen
canyon	cape	capital	carrot	cashier
casino	castle	catfish	cattle	cave
cavern	caviar	cavity	cayenne	celery
cellar	celling	cement	cereal	ceremony
chamber	channel	chapel	cherry	chestnut
chicken	chimney	church	cinema	city
clay	cliff	climate	clinic	closet
cloud	clover	club	coal	coast
coconut	coffee	college	colony	comet
company	compass	concert	congress	coon
copper	coral	corn	corridor	cottage
cotton	couch	country	county	coyote
crab	creek	crescent	crevice	cricket
crimson	crop	crystal	cucumber	daisy
dawn	daylight	debris	deck	deer
delta	desert	desk	device	diamond
diner	dirt	ditch	dogwood	dome
donkey	door	doorstep	doorway	dove
downhill	downpour	drawer	dresser	drizzle
drone	duck	dusk	eagle	earth
east	eclipse	edifice	elephant	elevator

ember	emerald	engine	estate	evening
facility	factory	farm	feather	fern
ferret	field	firefly	firewood	fish
flax	flea	flood	floor	flora
flower	fluid	foliage	food	foothill
footstep	forage	forenoon	forest	forestry
fortress	fossil	fountain	frog	frost
fruit	funeral	furnace	gadget	gala
gallery	garland	garlic	gate	gear
ginger	glass	globe	goat	goldfish
golf	gondola	goose	gorilla	grain
granite	grape	grass	gray	green
grocery	groove	ground	grove	growth
gulf	hail	Hall	halo	hardwood
hare	harvest	hawk	hazard	haze
helmet	hemp	herb	heritage	highland
highway	hill	honey	hoof	horse
hospital	hotel	hound	house	humidity
husk	iceberg	icicle	industry	inland
insect	iris	Iron	island	isle
ivory	jail	juice	jungle	kale
kernel	kingdom	kitchen	kitten	knob
lagoon	lake	lamb	lamp	land
landmark	lane	larva	latitude	lavatory
lavender	lawn	Leaf	lemon	leopard
lettuce	library	Lily	lime	lion
liquid	lizard	lobby	location	loft
lounge	lowland	lumber	luncheon	machine
mainland	mammoth	mansion	manual	manure
maple	marble	market	marsh	material
meadow	medicine	melon	metal	midday
midnight	milk	mill	mine	mineral
mist	monkey	moon	morgue	morning
mosquito	moss	moth	motor	mountain
mouse	movie	muck	mulberry	mule
mushroom	musical	nation	nature	navy
noon	north	nunnery	nursery	oatmeal
ocean	office	officer	olive	onion
opera	orange	orchard	orchid	outlet
oven	oyster	paddle	paint	palace
paper	parade	park	parsley	party
patio	pavement	pavilion	peach	peacock
peak	peanut	pear	pearl	pebble
pecan	pest	phone	piazza	picnic

pier	pigeon	pine	pinnacle	pipe
placard	plain	plane	planet	plant
plastic	plateau	platform	platinum	plaza
plough	plum	police	pollen	pond
pool	port	poster	postman	potato
province	public	puddle	pumpkin	puppy
purple	pyramid	rabbit	radiator	radish
rail	railing	railroad	railway	rain
rainbow	raindrop	raisin	rake	ranch
ranger	recital	redwood	reptile	resort
rhino	ripple	river	road	roadside
roadway	rock	roof	room	rose
saddle	salary	salmon	sapphire	sardine
scallop	scene	scenery	school	seashore
season	seaweed	seed	semester	senate
senator	sewer	shade	shark	sheep
shelf	shell	sheriff	ship	shop
shore	shrimp	sierra	sink	skunk
slope	snake	snow	soap	society
sofa	soil	south	speaker	spider
spinach	splash	spring	squirrel	stadium
stage	stair	stairway	star	station
statue	steel	stock	stone	store
storm	stove	stream	student	suburb
subway	suite	summer	summit	sundown
sunlight	sunrise	sunset	sunshine	surgery
swamp	swan	table	tablet	tavern
taxi	teacher	temple	tent	terrace
tide	tiger	Tile	timber	tire
toaster	toilet	Toll	tomato	tornado
torrent	tower	town	traffic	trail
train	tree	tribe	trolley	truck
tulip	tunnel	turkey	turtle	twig
twilight	universe	upland	utensil	vacuum
valley	valley	vault	villa	village
vine	vineyard	vista	volcano	vulture
wage	wagon	waiter	walnut	weapon
weather	west	whale	wheel	willow
wind	window	winter	wolf	wood
woodland	worker	workshop	world	yacht
yard	year			