Supplementary Materials for

The Order of Disorder: Deconstructing Visual Disorder and Its Effect on Rule-Breaking

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

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, Rand, & Zeckhauser, 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. (2014) showed that AMT participants passed an instructional manipulation check (Oppenheimer, Meyvis, & Davidenko, 2009) 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. Attrition rates

Table S1 presents the attrition rates by experiment.

	Condition	Started	Finished	Attrition %
Experiment 1		125	105	16%
Experiment 2	Scrambled-edges	105	95	10%
	Scrambled-colors	109	96	12%
Experiment 3		264	222	16%
Experiment 4	Visual order	280	206	26%
	Visual disorder	283	198	30%
Experiment 5	Visual order	251	198	21%
	Visual disorder	248	207	17%
Experiment 6	Visual order	223	197	12%
	Visual disorder	232	197	15%
Experiment 7	Visual order	55	49	11%
	Visual disorder	55	49	11%

Table S1. Attrition rates by experiment. The first column includes those who got to the consent form. The second column includes those who got to the demographics form.

3. Experiment 1 supplement

Table S2 presents the correlations between the low-level visual features and disorder ratings.

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

Table S2. 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, β = 0.53, $t(256) = 4.98$, $p < .001$, $\eta_p^2 = .088$, and asymmetry, β = 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.

4. Experiment 2 supplement

Table S3 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 S3. Experiment 2 correlation matrix. N = 260; *** $p < .001$

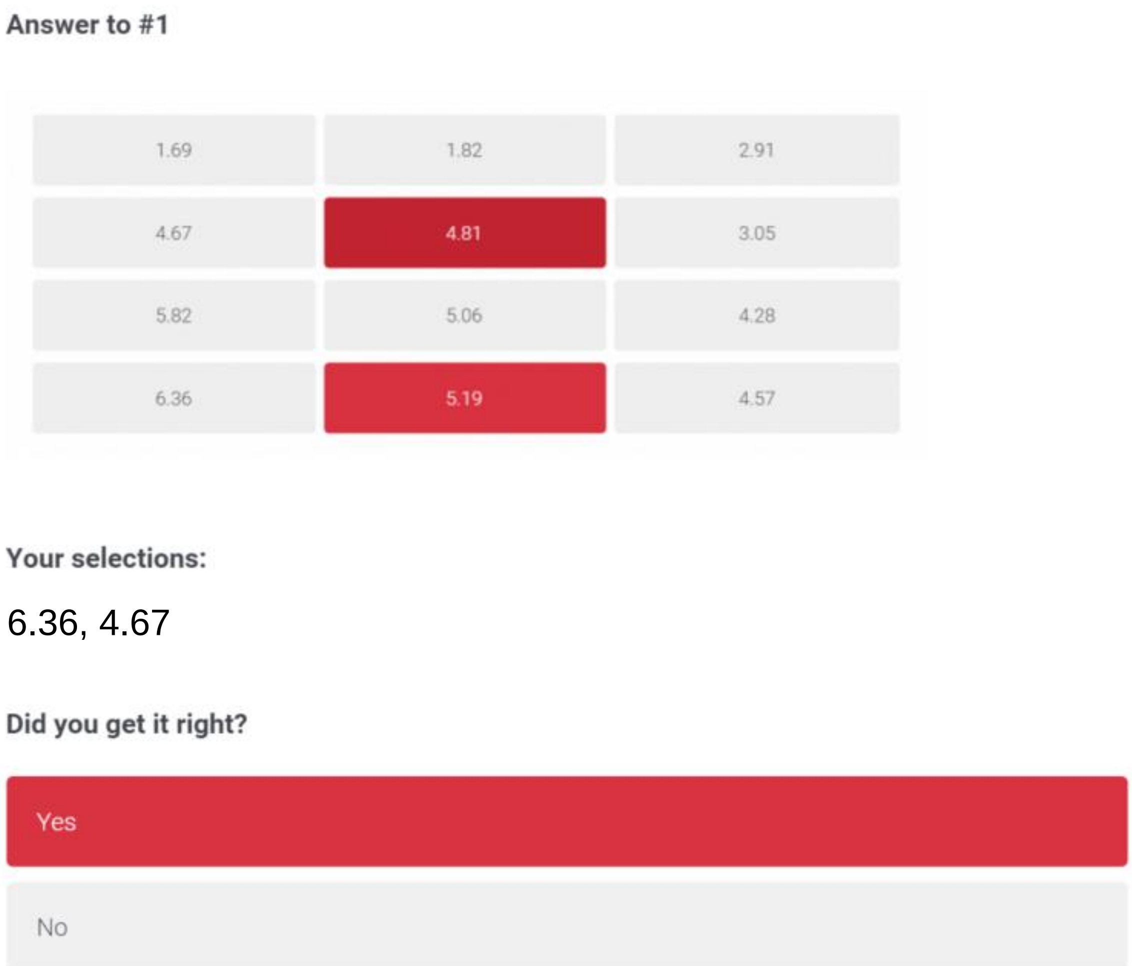
5. Experiment 3 supplement

Table S4 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 S4. Experiment 3 correlation matrix. N = 260; *** p < .001, ** p < .01

6. Screenshot showing how people cheated in Experiments 4-6



Answer to #1

1.69	1.82	2.91
4.67	4.81	3.05
5.82	5.06	4.28
6.36	5.19	4.57

Your selections:

6.36, 4.67

Did you get it right?

Yes

No

7. Experiment 4-6 supplement

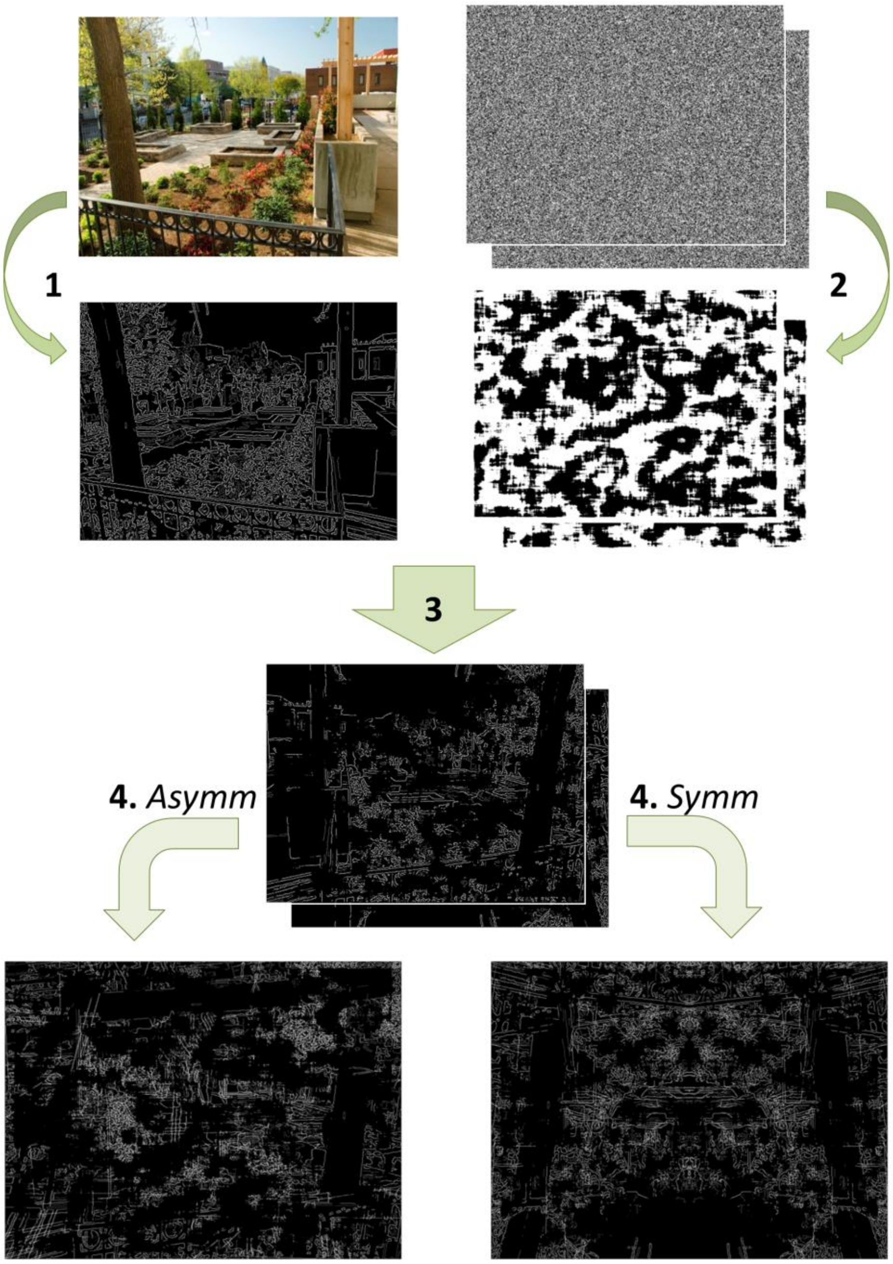
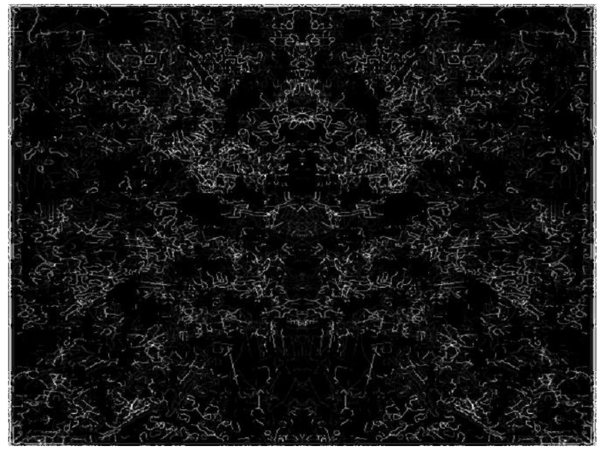
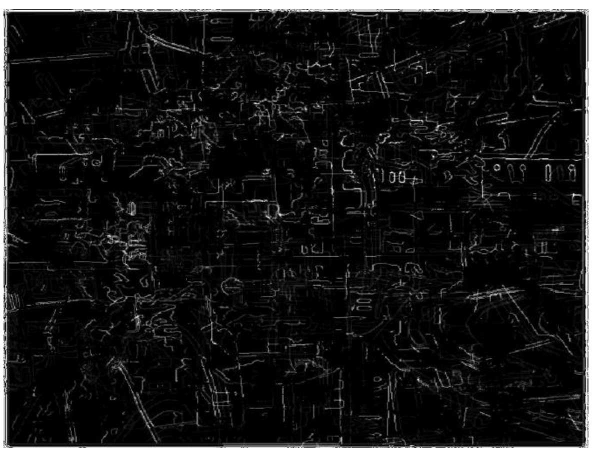


Figure S1. The process that results in symmetrical or asymmetrical scrambled edge stimuli, which depends on the two beginning 0 and 1 random matrices (top left) being identical (for symmetrical stimuli) or different (for asymmetrical stimuli). Process 1: The edge map of the picture is created. Process 2: Two random masks of patches each having, on average, half a surface of 1s and half a surface of 0s, are created by convolving a 30*40 median filter with two random 600*800 matrices each containing 1s and 0s. Process 3: The edge map is multiplied by each of the patch masks to lose half of its edges at random in two identical patch masks (for symmetrical stimuli) or two different patch masks (for asymmetrical stimuli) ways. Process 4: One of the output pictures from process 3 is flipped on the x-axis and overlaid on the other one. If the two patch masks are identical, this results in a symmetrical stimulus (bottom right), and if they are different this results in an asymmetrical stimulus (bottom left).

Symmetry + Visually Ordered Edges



Symmetry + Visually Disordered Edges



Asymmetry + Visually Ordered Edges

Asymmetry + Visually Disordered Edges

Figure S2. 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).

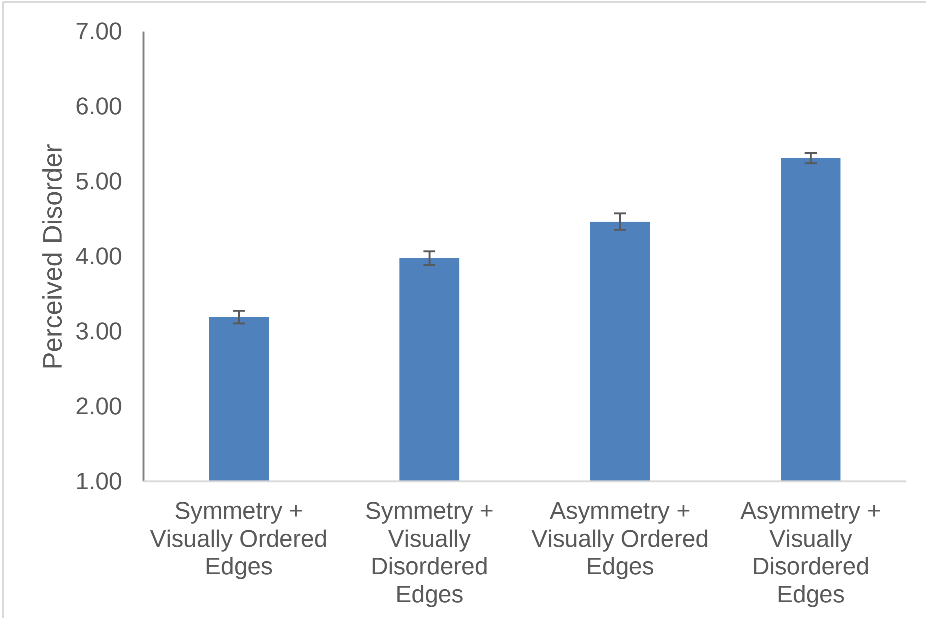


Figure S3. Mean disorder ratings for stimuli pretested for Experiment 4. The 30 most disorderly (all from asymmetry + visually disordered edges) and orderly (all from symmetry + visually ordered edges) stimuli were used for our manipulation of visual disorder vs. visual order in Experiments 4-7. Error bars indicate 95% CI of a one-sample t-test.

8. Results of preliminary experiment to Experiment 4 (lower cheating incentives)

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 ($M = 4.59$, $SD = 0.77$).

Cheating analysis: actual performance vs. reported performance. First we examined actual performance vs. reported performance in the visual-order vs. visual-disorder condition. 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. 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-effects 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, $t(401.00) = 11.48$, $p < .001$, 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, $t(401.00) = 0.44$, $p = .662$, and neither was the simple effect of visual condition within reported performance, $t(691.00) = 0.29$, $p = .769$.

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 relative 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, indicated that these extra cheaters tended to cheat by only a little bit.

References

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.

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.

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.

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.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872.

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.

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.