How big should this object be?
Perceptual influences on viewing-size preferences

Yi-Chia Chen (陳鴨嘉)[1,2], Arturo Deza[1,3], & Talia Konkle[1]

[1]Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, MA 02138, USA
[2]Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Box 951563, Los Angeles, CA 90095, USA
[3]Center for Brains, Minds and Machines, Massachusetts Institute of Technology, MIT Bldg 46-3160, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Running Head : Perceptual Influences on Size Preference

Addresses for correspondence : Yi-Chia Chen
Department of Psychology
University of California, Los Angeles
1285 Franz Hall
Box 951563
Los Angeles, CA 90095

Email : yichiachen@g.ucla.edu; deza@mit.edu; tkonkle@fas.harvard.edu

Abstract

When viewing objects depicted in a frame, observers prefer to view large objects like cars in larger sizes and smaller objects like cups in smaller sizes. That is, the visual size of an object that "looks best" is linked to its typical physical size in the world. Why is this the case? One intuitive possibility is that these preferences are driven by semantic knowledge: For example, when we recognize a sofa, we access our knowledge about its real-world size, and this influences what size we prefer to view the sofa within a frame. However, might visual processing play a role in this phenomenon—that is, do visual features that are related to big and small objects look better at big and small visual sizes, respectively, even when observers do not have explicit access to semantic knowledge about the objects? To test this possibility, we used "texform" images, which are synthesized versions of recognizable objects, which critically retain local perceptual texture and coarse form information, but are no longer explicitly recognizable. To test for visual size preferences, we used a two-interval forced choice task, in which each texform was presented at the preferred visual size of its corresponding original image, and a visual size slightly bigger or smaller. Observers consistently selected the texform presented at the canonical visual size as the more aesthetically pleasing one. These results suggest that the preferred visual size of an object depends not only on explicit knowledge of its real-world size, but also can be evoked by mid-level visual features that systematically covary with an object's real-world size.

Keywords

Aesthetic preferences; Size; Mid-level visual features; Texture; Object recognition

1. Introduction

One of the most frequent everyday activities we engage in is inspecting objects. When we detect a bird in a tree, find a box of snacks lying deep in the fridge, or spot a product in an aisle of a shopping mall, we gather more information about the object by getting closer to it and stopping at a proper distance to look at it. This idea that each object has an optimal viewing distance, and that perception draws us to move our bodies to the distance that balances between deficiency on one hand (too far away) and excess on the other (too close), has been highlighted by philosophers of perception (Merleau-Ponty, 1962; Kelly, 2010). By moving closer to or farther from an object, the observer can adjust the visual size that the object subtends in their visual field. Indeed, research has found that given a picture of an object, there is a systematic, or "canonical", visual size at which the object "looks best" in, and, curiously, this visual size is linked to the physical size of the object: When viewing items with a bigger physical size (e.g., a car), we prefer to view them at a bigger visual size; and when viewing items with a smaller physical size (e.g., a cup), we prefer a smaller visual size (Konkle & Oliva, 2011; Linsen, Leyssen, Sammartino, & Palmer, 2011; see also Eckstein, Koehler, Welbourne, & Akbas, 2017).

1.1 Knowledge of physical sizes

What is the nature of these physical size representations that drive the systematic canonical visual sizes? A likely candidate is the rich real-world size knowledge we eagerly pick up as we experience the world, evident in toddlers and even infants (e.g., Granrud, Haake, & Yonas, 1985; Long, Moher, Carey, & Konkle, 2019a, 2019b; Sensoy, Culham, & Schwarzer, 2020; Yonas, Pettersen, & Granrud, 1982). We can clearly learn the physical sizes of objects from our

own past sensory experience (e.g., the size of our favorite toy from childhood) and build this knowledge further by incorporating semantic knowledge (e.g., even if you have never seen a "ranchu" or a picture of one, if you learn that it is a kind of goldfish, you might then infer that it is roughly the size of a typical pet goldfish; see Chen, Lu, & Holyoak, 2014). Further, size knowledge can be completely abstracted from direct sensory experience; for example, we can represent and reason about the physical size of an atom, the earth, or even of a unicorn. Interestingly, this knowledge of objects' real-world sizes seems to influence how we spatially allocate our visual attention (Collegio, Nah, Scotti, & Shomstein, 2019), demonstrating an example of interaction between size knowledge and other aspects of cognition. Thus, one possible account of canonical visual size is that it arises as a consequence of our abstract physical size knowledge.

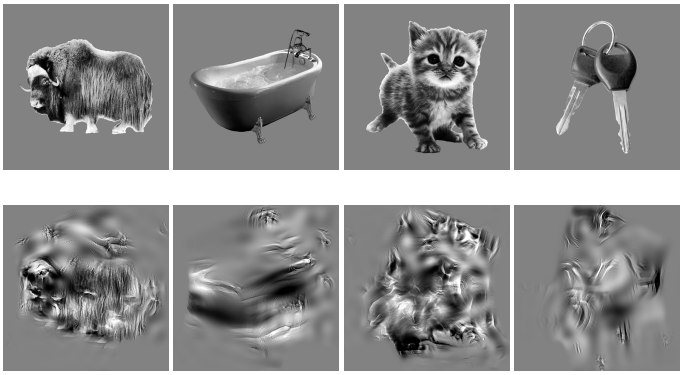
1.2 Perception of physical sizes

Interestingly, along with the rich size knowledge we have, our visual systems seem to maintain perceptual representations that distinguish objects of different physical sizes as well. This point is revealed behaviorally with several different methods: Visually searching for a picture of a big object (e.g., a building) among an array of pictures of small objects (e.g., a flashlight, a cap, etc.) is faster than when searching for the same picture of the big object among other pictures of big objects (e.g., a bed, a boat, etc.; note that the visual size of all the items in the array is the same; Long, Konkle, Cohen, & Alveraz, 2016; Long et al., 2019a). This result indicates that there are systematic perceptual differences between big and small objects (as classes), that can be used to speed up visual search processes. For example, bigger objects tend to be boxier with higher spatial frequency, small objects tend to be curvier and smoother (Konkle, 2011; Long et al., 2016).

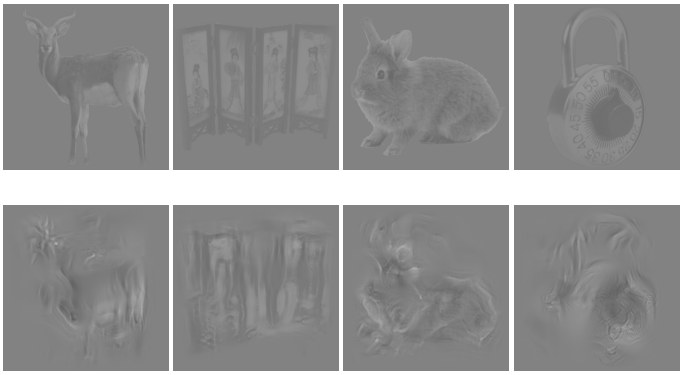
An even stronger case for these systematic *perceptual* differences among objects of different physical sizes comes from a line of work using "texform images"—these are distorted images synthesized from images of recognizable objects, which critically retain some perceptual texture and coarse form information, while "knocking out" the object identity (See Figure 1; Long et al., 2016; Deza, Chen, Long, & Konkle, 2019). The facilitated visual search for big among small objects (and vice versa) persists with texform images (Long et al., 2016): that is, texforms of big objects were faster to find among texforms of small objects than among texforms of big objects. These results further support the claim that there are systematic perceptual differences in the shape and texture of objects of different real-world sizes.

These systematic perceptual differences between big and small objects not only influence visual search but also are powerful enough to interfere with even simple perceptual judgments about what image is bigger or smaller *on the screen*, a task that does not require any access to the identity or the real-world size of the objects. That is, people are faster to select the visually smaller of two objects on the screen if it is in fact smaller in the real world (Konkle & Oliva, 2011). Critically, the same effect was found using texforms (Long & Konkle, 2017; Long et al., 2019b): For example, people were faster to pick the visually smaller of two unrecognizable texforms, when the visually smaller texform was generated from a small object (e.g., key) than when the visually smaller texform was generated from a big object (e.g., piano). Thus, perceptual feature differences between big and small objects are sufficient to automatically influence visual size judgments. Finally, complementing these behavioral signatures, there is also evidence for different ventral stream sensitivity to these perceptual features: Different regions of cortex respond more to big object texforms than small object texforms (and vice versa), with highly similar large-scale ventral stream topography as evoked when viewing recognizable objects with big vs small real-world sizes (Long, Yu, & Konkle, 2018).

Experiment 1 & 2a



Experiment 2b



Taken together, these studies prompt another possible account for canonical visual size: The preferred visual size of an object may arise as a consequence of perceptual processing (rather than explicit recognition and reasoning), where certain kinds of visual features (e.g., curvier features evident in smaller objects) are processed more effectively in certain visual sizes (e.g., relatively smaller) than other visual features (e.g., more rectilinear edge information present in bigger objects). As such, it is possible that an underlying cause of the systematic canonical visual sizes are in fact perceptual in nature. Our goal in this study is to explore this possibility.

1.3 The current study: Perceptual contributions to canonical size?

Here, we tested if the canonical visual size of objects can be observed even when the images of objects have been "texformed" so they are no longer recognizable, preventing explicit access to the objects' identities and associated real-world size knowledge. We first asked in Experiment 1 if there is systematic canonical visual size for texform images, using a method of adjustment, which, to foreshadow, yielded equivocal results. We then turned to a forced-choice paradigm in Experiment 2a and its replication Experiment 2b, which showed clear and replicable results.

2. Experiment 1: Method of Adjustment

In the first experiment, we examined whether intact and texform images have systematic canonical visual sizes using a method of adjustment task: Subjects were asked to rescale an image presented on the screen until it "looks best". The key questions are: First, do we replicate Konkle and Oliva (2011), showing consistent preferred visual sizes for intact recognizable

objects related to their real-world size? And, second, do texforms show consistent preferred visual sizes, related to real-world size, corresponding to the original images?

2.1 Method

2.1.1 Participants

Fifteen naive subjects (6 females, 8 males, and 1 other gender; all with normal or corrected-to-normal visual acuity) from the Harvard University community completed individual 60-min sessions in exchange for a small monetary payment or a course credit. This sample size was preregistered[1] and was fixed to be identical across all experiments reported here. Four subjects were replaced based on predetermined exclusion criteria reported in Section 2.1.5 Exclusions.

2.1.2 Apparatus

The experiment was conducted with custom software written in Python with the PsychoPy libraries (Peirce, Gray, Simpson, MacAskill, Höchenberger, Sogo, et al., 2019). The subjects sat approximately 60 cm without restraint from an iMac computer (with a viewport of 47.6 cm x 26.7 cm and effective resolution of 2048px x 1152px).

2.1.3 Stimuli

The final stimulus set consisted of 40 original recognizable images and 40 corresponding texforms images (depicting 10 big animals, 10 big objects, 10 small animals, and 10 small objects). To generate this curated and controlled set of images, we used the following procedure.

---

[1] For preregistration of Experiment 1, visit https://aspredicted.org/at3v7.pdf . The only deviation of experiment details from the preregistration is that the block order was not counter-balanced but alternated before subject exclusions.

First, a superset of 180 recognizable images were collected from various sources including stimuli from previous works (Long et al., 2018; Konkle & Caramazza, 2013; Konkle & Oliva, 2012) and Google images—consisting of 90 big items (big enough to support an adult human being) and 90 small items (small enough to be held by one hand), with an equal balance of animals and man-made objects. These images went through preprocessing to (a) remove the backgrounds, (b) crop to the smallest square that envelops the items, (c) resize to 512px x 512px, (d) convert to grayscale, (e) equalize their luminance and luminance histograms, and (f) place in the center of a gray background of 640px x 640px. The details of the preprocessing can be found in Appendix A. The resulting images are referred to as intact images (see Figure 1).

Next, the corresponding texform images (see Figure 1) were generated from these intact images following the method detailed in Deza et al. (2019), which is a variation and extension of the method used in Long et al. (2018). To overview, each texform image was synthesized from random noise image seed, coerced to match the first and second order image statistics of each intact input image (following Freeman & Simoncelli, 2011). The size of the pooling windows over which these textural image statistics were computed reflects a peripheral placement in a simulated visual field (i.e., with small enough pooling windows with respect to the visual size of the depicted object to retain some coarse form information, but large enough with respect to the visual size of the depicted object to texturize the content, usually beyond recognition). Note this slightly modified texform algorithm enabled us to synthesize higher resolution texform images (640 px x 640 px) than in Long et al. (2018) (180 px x 180 px), for more on the method see Deza et al. (2019).

Finally, following the generation of these candidate texform images, we conducted an online pretest to test for texform recognizability (for details of the pretest, see Appendix B). Based on these results, we selected a final set of 40 pairs of intact and texform images (20 big

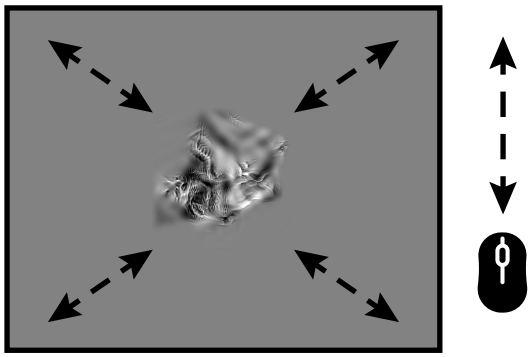
items and 20 small items, half depicting animals and half depicting inanimate objects), where the texform images were unrecognizable at the basic level for at least 15 out of 18 pretest observers. (This was still a relatively low cut-off, so our experiment included a recognition post-test and excluded for each subject the images they recognized from the analysis.)

2.1.4 Procedure and Design

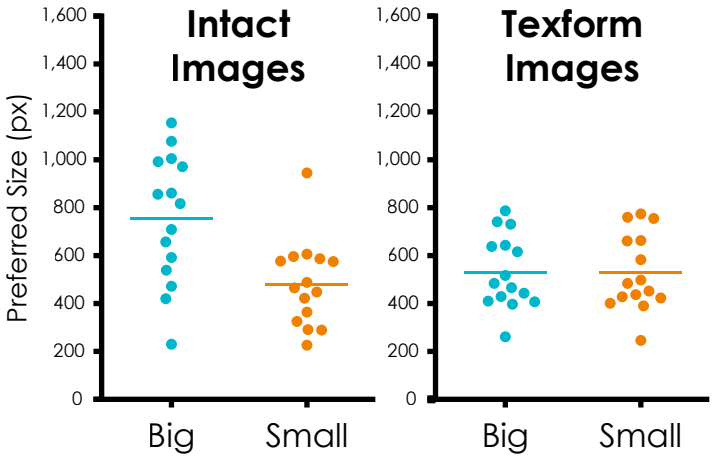
Each trial began with a 400 ms blank gray screen (matching the background color of all the images) followed by the presentation of a single centered image. The subjects were instructed to move the invisible cursor up and down to make size adjustments to the image. Moving the cursor up increased the visual size and moving it down decreased the visual size (see Figure 2a). The allowed size ranged from 5px x 5px to 1552px x 1552px. All images were initially presented at the medium size of 778px x 778px. The subjects made as many adjustments as they liked, to make the image the size they found "most visually pleasing". They then clicked the mouse to submit their responses. (Mouse clicks within 300 ms of the onset of the images were recorded but ignored.)

In each "intact block", all 40 intact images (2 real-world sizes x 20 items) were each presented once in a block in randomized order; and in each "texform block" the 40 texform images were presented in a randomized order. Five subjects completed 4 intact blocks followed by 4 texform blocks, and 10 subjects completed 4 texform blocks followed by 4 intact blocks. All completed a total of 320 trials. The block order alternated between subject before subject exclusions (data from all but 1 subject showed the same pattern in the critical analysis, regardless of the block order). Subjects took 3 self-paced breaks when they completed 25%, 50%, and 75% of the experiment. Before the main experimental trials, subjects completed 4 practice trials with images (1 big object, 1 big animal, 1 small object, 1 small animal, either all intact or all

A. Method of Adjustment



B. Experiment 1 Results



texform, depending on the first block type); these images never appeared in the main experimental trials. The subjects were not told about the nature of the texform images.

After the adjustment task, subjects completed a recognition test on all texforms to assess whether these specific subjects recognized the texform images (though note that these subjects also saw the corresponding intact images in the same setting). They were told that the texform images they saw were "made from images of objects by distorting the images while keeping their textures". They then viewed all texform images one by one again and typed in with a keyboard what they thought was depicted in each image.

2.1.5 Exclusions

The responses from the recognition test were graded by the first author before looking at the adjustment data: To be conservative at estimating the unrecognizability of texform images, any response that named an object with a similar size and shape from the depicted object was considered correct. Any adjustment trials from stimuli with their texform version recognized in the recognition test were discarded, along with adjustment trials with mouse clicks within 300 ms of the onset of the images. Three subjects had more than 20% of trials discarded and thus were replaced with new subjects.

Next, we tested the consistency of the preferred visual sizes of the intact images, since this is a necessary precondition for examining textform feature contributions to this preferred visual size. To estimate the reliability of the preference, we computed the correlation between the selected sizes across the first half and the last half of trials for the intact, recognizable images. One subject was removed based on having low reliability (r'<.5), and was replaced, yielding a final average reliability of the preferred visual sizes for intact images of r'=.82 (SD=.15). Unlike the preregistered plan, we did not exclude subjects based on the reliability of the preferred visual size of texforms because the reliabilities were generally very low (r'=.21, SD=.25). After

subject replacements, the mean recognition rate was 7.3% (*SD*=5.9%) and a total of 68 out of 4800 trials were discarded due to early mouse clicks.

2.2 Results and discussion

First, we examined whether, for intact recognizable objects, subjects chose consistent preferred visual sizes that were related to the real-world size of the depicted object. The results are shown in Figure 2. Overall, we found that people did show the signature preferences. Subjects preferred to view the big items at a bigger visual size (757px, *SD*=268px) compared to small items (480px, *SD*=180px; $t(14)=3.78$, $p=.002$; 14 out of 15 subjects, $p=.002$), regardless of whether the images depicted animals or inanimate objects, $F(1,14)=3.45$, $p=.085$). Thus, these results are consistent with the canonical visual size effect (Konkle & Oliva, 2011).

On the other hand, with the texform images, the reliability of the preferred sizes was quite low ($r'=.21$, *SD*=.25), indicating subjects didn't select similar visual sizes across repeat presentations of the same texform image. Further, we did not observe the signature preference where the big items were preferred at bigger visual sizes than small items (531 px, *SD*=153, vs 530 px, *SD*=160; $t(14)=0.07$, $p=.944$; 10 out of 15 subjects, $p=.302$). Thus, the simple act of resizing until the texform image "looks best" did not yield consistent preferred visual sizes.

While these results could indicate an actual lack of canonical size for texform images, the unreliable responses may also be related to the nature of the adjustment task. For example, in facing the unfamiliar texform images, it is possible that subjects felt less confident in making a choice from the unlimited options given by an adjudgment task, leading to a family of unconstrained strategies. We thus performed Experiment 2a and 2b with a more rigorous psychophysical method to probe for the existence of visual size preferences in texform images. Additionally, the lack of effect in this adjustment task has one interpretive benefit—that is, it

provides further support that subjects are not systematically recognizing these texform images as something (if they were, the sizes would be consistent across repetitions).

3. Experiment 2: Forced-Choice Task

Experiment 2a and 2b cut the number of preferred visual size options down from unlimited to only two, using a forced choice paradigm. That is, subjects could toggle between two options and selected the one that looked best. We conducted two versions of this experiment: In the first version, we created a larger stimulus set drawn from the same superset as reported in the experiment above. In the second version, intended as a replication experiment with some generalization, we changed the stimulus set again, in order to dovetail more closely with previous work, using a subset of the original texform images used by others (Long et al., 2016; Long & Konkle, 2017; Long et al., 2018; Wang et al, in prep; Grootswager et al. 2019; see Figure 1). These two versions of the experiment (Experiment 2a and 2b) were otherwise identical, except for the stimuli used.

3.1 Method

The experimental apparatus and general procedures were similar to Experiment 1, except as noted here.

3.1.1 Participants

Each experiment was completed by 15 naive subjects. (Experiment 2a: 10 females, 5 males; Experiment 2b: 6 females, 9 males). Subjects were replaced based on preregistered[2]

[2] For preregistration of Experiment 2a, visit https://aspredicted.org/pf87y.pdf . The only deviation of experiment details from the preregistration is that the block order was not counter-balanced but alternated before subject exclusions.

exclusion criteria reported in Section 3.1.4 Exclusions (6 excluded and replaced in Experiment 2a; 1 excluded and replaced in Experiment 2b).

3.1.2 Stimuli

In Experiment 2a, 50 pairs of intact and texform images were repicked from the superset of 180 processed images as described in Experiment 1. Half of the images depicted big items and half small items (with a balanced selection of animals and inanimate objects). Based on a pilot study using the same adjustment task from Experiment 1, these images were selected to maximize the range of canonical visual sizes, while also maintaining a generally balanced set across real-world size and animacy (12 big and 13 small animals; 12 big and 13 small objects). The images were then scaled down to 440px x 440px, which is a lower resolution than in Experiment 1 (based on pilot studies, this design choice helped to ensure that the preferred sizes of intact images were well within the size of the screen).

In Experiment 2b, 50 pairs of images were selected from the stimuli from Long, et al. (2018), available online (https://konklab.fas.harvard.edu/). The main difference of these texforms is that they have a lower spatial resolution.[3] The images were selected to include 25 animals and 25 objects and maximize the range of canonical visual sizes based on a pretest, this resulted in 19 big items (8 big animals, 11 big objects) and 31 small items (17 small animals, and 14 small objects). Note that, here the division of items into "big" and "small" is less relevant, as we can treat size here as a continuous variable.

3.1.3 Procedure and Design

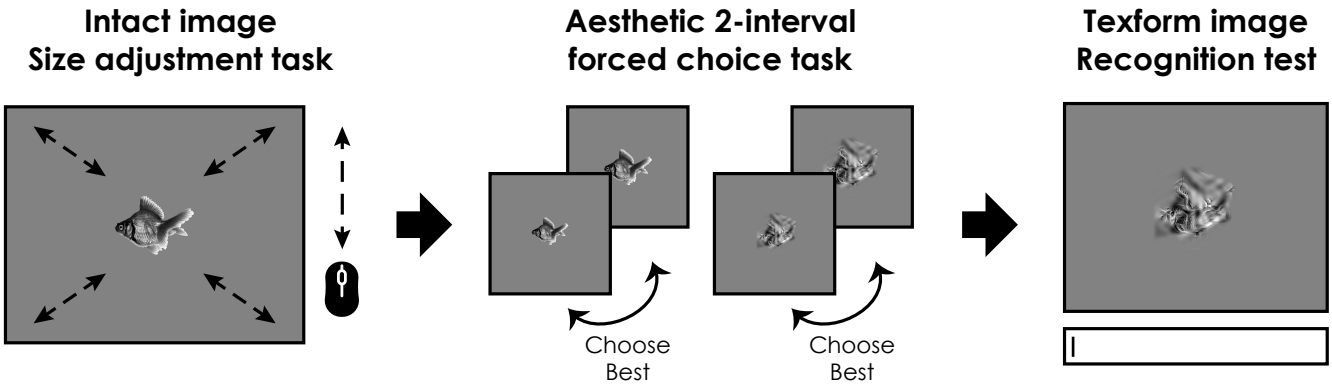
[3] In their generation procedure the intact images were first scaled down to 180px x 180px and embedded in a 640px x 640px gray background. This image served as the input to generate the texform. After the synthesis, 192px x 192px area centered at where the input images were embedded was cropped and rescaled back to 440px x 440px. Finally, the four edges were blurred so that they gradually faded into their backgrounds.

The subjects completed 3 tasks in order: (a) an adjustment task on intact images, to obtain a canonical visual size estimate for each item, (b) the main forced choice task, to select which of two visual sizes of the same image was more aesthetically pleasing, completed for both intact and texform images in different blocks, and (c) a post-test assessing recognition on the texform images. This procedure is depicted in Figure 3a.

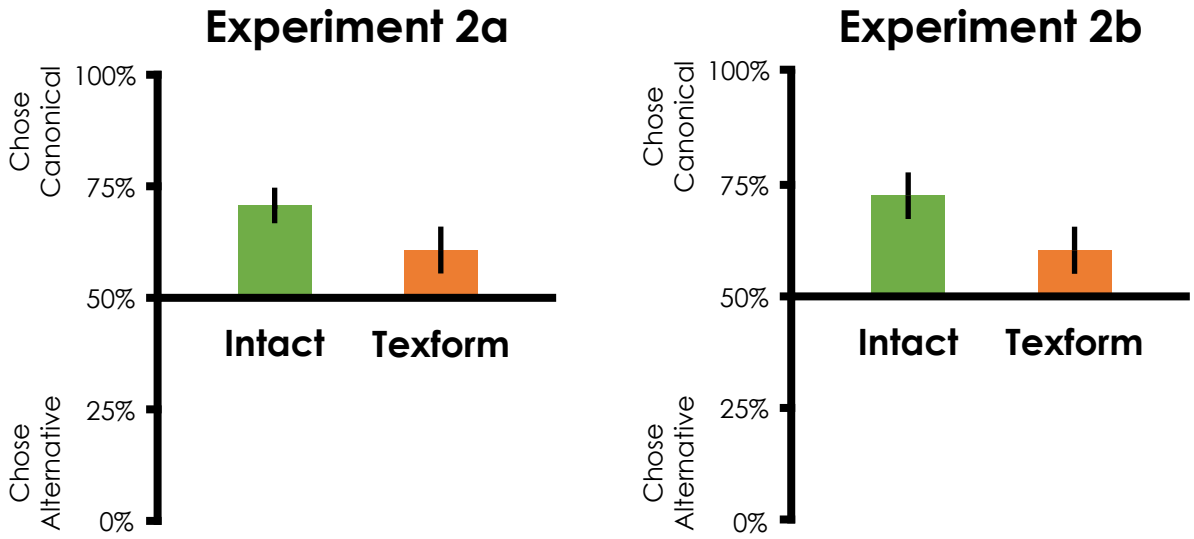
Size adjustment task on intact images. As in Experiment 1, subjects moved the mouse to adjust the visual size of an image on the screen, and clicked when the image "looked best." This task was identical to the adjustment task in Experiment 1, except that it consisted of only 4 intact blocks of 50 trials, the allowed size ranged from 20px x 20px to 2152px x 2152px, with images initially presented in 1086px x 1086px, the click detection started earlier at 100 ms after the onset of the images, and there was only 1 practice trial. For each item, its canonical visual size for that subject was calculated by averaging the selected sizes from the repetitions, after excluding trials with a response time (RT) less than 300 ms and excluding items that had more than half of the 4 trials excluded. Only the items that yielded canonical visual size smaller than 871px x 871px (so that the images stayed well within the monitor's size) entered the next task (with both the intact images and their corresponding texform images). Critically, these canonical visual sizes were used to set the choice options for the next task.

Aesthetic 2-interval forced choice task. On each trial, subjects viewed a single image and toggled between two sizes with a key press. They were asked to toggle to view both sizes as many times as they liked and decide which of the sizes "looks more aesthetically pleasing". Unbeknownst to the subjects, one of the size options was the average canonical visual size they picked for that item in the adjustment task, and the other was 30% different in diagonal length (either visually bigger, or visually smaller, balanced across trials; see Figure 3a for example displays from Experiment 2a and 2b). Which size option was shown first was randomized.

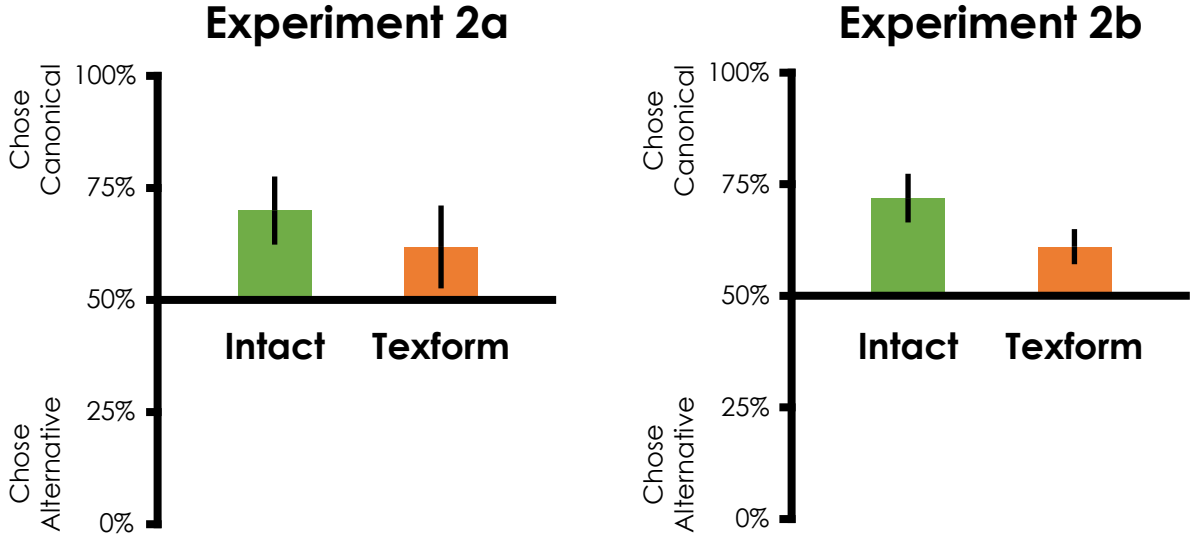
A. Procedure



B. Results from unrecognized items



C. Results from recognized items



Subjects performed this task in separate blocks for intact and texform images. Critically, in the texform block, the visual sizes were based on the canonical visual sizes of the corresponding intact images. For both intact and texform blocks, we calculated the percentage of trials in which subjects picked the canonical visual size options as the key outcome measure. The block order alternated between subjects before subject exclusions (resulting in 9 subjects completing the texform block first, and 6 completing intact block first in Experiment 2a; with 7 completing the texform block first and 8 completing the intact block first in Experiment 2b).

Texform image recognition test. Finally, as subjects received extensive exposure to the intact images as well as the corresponding texforms, we next tested the recognizability of the texforms that were included in the forced choice task. The same procedure and grading as in Experiment 1 were performed, and the average texform image recognition rate was 24% (*SD*=14%) for Experiment 2a and 56% (*SD*=12%) for Experiment 2b. Below we report analyses from unrecognized and recognized items separately.

3.1.4 Exclusions

The following preregistered exclusion criteria were applied: (a) forced choice trials with RT less than 300 ms, (b) forced-choice trial without any toggling (i.e., the subject picked the first option without viewing the second option), (c) subjects with more than or equal to 5% trials excluded in either the adjustment or the forced choice task, (d) subjects who had less than 12 items (i.e., 12 intact and 12 texform images)[4] entered into the forced choice task, and (e) subjects who had by-item split-half reliability lower than 0.5 in the adjustment task. As noted above, excluded subjects were replaced to achieve the pre-registered N=15 for each experiment.

[4] The preregistration specified 30 items as the criterion; however, this ended up being too strict and excluded most of the subjects. We thus decided on 12 items in Experiment 2a and replicated the results in Experiment 2b with this new criterion.

3.2 Results and discussion

3.2.1 Main analyses with unrecognized items

We first analyzed the visual preference data from the forced choice task, but only including trials in which the texforms were not subsequently recognized during the recognition test. (This included 76.9% of the forced-choice trials in Experiment 2a, 43.9% in Experiment 2b.) The percent of trials in which subjects chose the canonical visual size rather than modulated size was plotted for both intact and texform blocks in Figure 3b.

Inspection of the figure reveals two main results, present in both experiments. First, subjects chose the canonical size over the alternative above chance, for intact images and, critically, also for texform images. Second, the canonical visual size preference was stronger in intact than in texform images. Indeed, one-sample t-tests confirmed all blocks were above the chance level of 50% (Experiment 2a: intact block, 71%, SD=7%, $t(14)$=10.90, p<.001, 15 out of 15 subjects, p<.001; texform block, 61%, SD=9%, $t(14)$=4.38, p=.001, 13 out of 15 subjects, p=.007; Experiment 2b: intact block, 73%, SD=9%, $t(14)$=9.24, p<.001, 14 out of 15 subjects, p=.001; texform block, 60%, SD=10%, $t(14)$=4.19, p=.001, 12 out of 15 subjects, p=.035), and the differences between intact and texform blocks were also significant (Experiment 2a: $t(14)$=3.16, p=.007, 13 out of 15 subjects, p=.007; Experiment 2b: $t(14)$=3.81, p=.002, 12 out of 15 subjects, p=.035).

These findings confirmed that subjects preferred to view texform images in the canonical visual size of their original versions compared to bigger or smaller alternatives. Thus, visual size preference persists partially after object recognition is disrupted, providing evidence that mid-level visual features contribute to the phenomenon of canonical visual size. However,

at the same time, the visual size preference was still stronger in intact images. The following exploratory analyses with recognized items help shed light on the interpretation of this effect.

3.2.1 Exploratory analyses with recognized items

The same analysis was performed on trials with recognized items (see Figure 3c). Inspection of the figure suggested the same three trends in both experiments: First, subjects chose canonical size over the alternative above chance level with both intact and texform images. Second, canonical size preference was stronger in intact than in texform images. Third, the data patterns were almost identical to the unrecognized trials. One-sample t-tests confirmed the first impression that all blocks was above the chance level of 50% (Experiment 2a: intact block, 70%, SD=14%, $t(14)$=5.63, p<.001, 12 out of 15 subjects, p=.035; texform block, 62%, SD=17%, $t(14)$=2.75, p=.016, 10 out of 15 subjects, p=.302; Experiment 2b: intact block, 72%, SD=10%, $t(14)$=8.60, p<.001, 15 out of 15 subjects, p<.001; texform block, 61%, SD=7%, $t(14)$=5.99, p<.001, 13 out of 15 subjects, p=.007). There was also a difference between intact and texform blocks, but this difference was only significant in Experiment 2b (Experiment 2a: $t(14)$=1.37, p=.193, 8 out of 15 subjects, p>.999; Experiment 2b: $t(14)$=3.07, p=.008, 11 out of 15 subjects, p=.118). This is likely due to the lower recognition rate in Experiment 2a than in 2b (24% vs. 56%), leading to fewer trials entering the analysis and thus higher variances (SD_{diff}=23% vs. SD_{diff}=12%).

The almost identical patterns found in unrecognized and recognized images suggests that explicit identity and/or size knowledge access is not a major factor in the effects found here. Texform images still showed weaker canonical visual size preferences than intact objects, even when the texforms were subsequently recognized (and thus could have potentially allowed access to physical size knowledge during the aesthetic choice task). These results provide

additional support that these consistent visual size preferences for texform images are driven by their visual features.

4. General Discussion

Our minds have at least two sources of information when it comes to representing the physical size of objects in the world: We can access knowledge about the objects' size attributes from knowing what they are (e.g., Chen et al, 2014), and we also perceive visual feature differences between objects of different sizes (e.g., Long et al., 2016). Here, we asked which kind of information is driving the systematic visual size preference, where we like to view big things big and small things small (Konkle & Oliva, 2011; Linsen et al., 2011). In three experiments, we first replicated the systematic visual size preferences for recognizable objects and found some evidence for the role of perceptual features in such preferences: While resizing texforms until they looked best did not show strong canonical visual sizes (Exp 1), we did find consistent preferences when given only two options (replicated across Exp 2a and 2b). These results demonstrate that visual size preferences, instead of only stemming from knowledge of objects' physical sizes, can also be evoked by the visual features that are preserved in unrecognizable texforms.

4.1 Mid-level visual features: Curviness?

Since the visual size preferences for texforms are systematic without object recognition, the information about the objects' physical sizes must be coming from the visual features. What kind of visual features carry the information about an object's physical size? While our experiments do not have direct evidence to pinpoint the responsible features, the use of texform images constrained the possibilities: Among visual features preserved in texform images, the

curviness of an object was found to correlate with the object's physical size (Konkle, 2011; Long et al, 2016). Curviness is also related to real-world size in large-scale neural organization (Long et al., 2018; see also Srihasam, Vincent, & Livingston, 2014; Yue, Robert, & Ungerleider, 2020), and has also been linked to eccentricity computations (e.g., see Ponce, Hartmann, Livingstone, 2017). Thus, these properties make curviness a candidate feature that perceptually contributes to aesthetic preference on visual size.

4.2 Size knowledge's role in aesthetics?

While we showed that perceptual processes contribute to canonical visual sizes, it is nevertheless likely that knowledge of physical sizes still plays a role in size preference in other contexts. In fact, telling people that objects they were viewing were "toys" (thus were physically small) reduced the canonical visual sizes by more than 50% (Konkle & Oliva, 2011, Experiment 4). Size knowledge also influences aesthetic experience in a very different way—through expectations and pleasant surprises. Famously, artists (e.g., Claes Oldenburg) created humongous statues of everyday objects, which induced aesthetic experience presumably through challenging our expectations (e.g., Van de Cruys & Wagemans, 2011). These kinds of aesthetic experiences have been argued to differ in intensity (and maybe in nature as well) from those that we may rely on to pick out a canonical visual size that simply "looks good" (e.g., Makin, 2017; Brielmann & Pelli, 2017).

4.3 The function of canonical visual sizes?

Why do we have canonical visual sizes? Of course, our study does not provide a direct answer but merely inspired two speculative ideas. One possibility is that it is simply a byproduct of ontogenetic and/or phylogenetic developments of visual systems: The visual

features' correlation with physical sizes gives rise to correlation with visual sizes in experience as well. For example, if we tend to see small objects in smaller visual sizes, and small objects tend to be curvier, our visual systems may process visually small curvatures more fluently than visually big curvatures. And this perceptual fluency leads to a more positive experience when viewing physically small objects in small visual sizes (as in Reber, Schwarz, & Winkielman, 2004). In this way, the size preference itself may not have a particular function but is just an indication of the visual system's tuning for features it commonly encounters.

Another possibility is that canonical visual size is in fact functional: It guides us to seek the proper viewing distances for each object to maximize the information intake and minimize the danger associated with getting close to unknown objects in the environment. For example, if we are looking at an unknown object, our knowledge cannot guide us to interact with it in a proper distance, but our visual systems may use heuristics based on visual features to induce aesthetic experience, which in turns motivate us to seek a proper viewing distance. This way, the function of visual size preference is to assist active learning by motivating us to modulate the visual inputs themselves, adding support to the idea that aesthetic experience interacts with perception (e.g., Chen & Scholl, 2014) and serves adaptive functions (e.g., Bar & Neta, 2006; Chen, Colombatto, & Scholl, 2018; Forman, Chen, Scholl, & Alvarez, submitted; Orians & Heerwagen, 1992). Avenues that examine the relevance and functional role of these systematic visual size preferences are open for future study.

Acknowledgement

For helpful conversation, we thank Bria Long, Brian Scholl, Felix Chang, Justin Halberda, and the members of the Harvard Vision Sciences Laboratory. This work was supported by an NSF CAREER BCS-1942438.

Author Contribution

Y.-C. Chen and T. Konkle designed the research and wrote the manuscript with input from A. Deza. Y.-C. Chen and A. Deza created the stimuli. Y.-C. Chen conducted the experiments and analyzed the data with input from T. Konkle.

Open Practice Statement

The supplementary file available online with this paper contains the preregistration, materials, and raw data for each experiment, which can be found at https://osf.io/pqvsr/?view_only=c36c5cbad60f4228a37230590db4fde3.

References

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

Brielmann, A. A., & Pelli, D. G. (2017). Beauty requires thought. *Current Biology, 27,* 1506–1513.

Chen, D., Lu, H., & Holyoak, K. J. (2014). The discovery and comparison of symbolic magnitudes. *Cognitive Psychology, 71,* 27–54.

Collegio, A. J., Nah, J. C., Scotti, P. S., & Shomstein, S. (2019). Attention scales according to inferred real-world object size. *Nature Human Behaviour*, *3*, 40–47.

Deza, A., Chen, Y. -C, Long, B., & Konkle, T. (2019, September 15). Accelerated texforms: Alternative methods for generating unrecognizable object images with preserved mid-level features. [Paper presentation]. *Conference on Cognitive Computational Neuroscience*, Berlin, Germany, https://ccneuro.org/2019/proceedings/0000879.pdf

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology, 27,* 2827–2832.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature neuroscience, 14,* 1195–1201.

Granrud, C. E., Haake, R. J., & Yonas, A. (1985). Infants' sensitivity to familiar size: The effect of memory on spatial perception. *Perception & psychophysics*, *37,* 459–466.

Grootswagers, T., Robinson, A. K., Shatek, S. M., & Carlson, T. A. (2019). Untangling featural and conceptual object representations. *NeuroImage, 202*:116083, 1–9.

Kelly, S. D. (2010). The normative nature of perceptual experience. In B. Nanay (Eds.), *Perceiving the world* (pp. 146–159). New York: Oxford University Press.

Konkle, T., & Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience, 33*, 10235–10242.

Konkle, T., & Oliva, A. (2012). A Familiar Size Stroop Effect: Real-world size is an automatic property of object representation. *Journal of Experimental Psychology: Human Perception & Performance*, *38*, 561–569.

Konkle, T. & Oliva, A. (2011). Canonical visual size for real-world objects. *Journal of Experimental Psychology: Human Perception & Performance*, *37*, 23–37.

Linsen, S., Leyssen, M. H., Sammartino, J., & Palmer, S. E. (2011). Aesthetic preferences in the size of images of real-world objects. *Perception, 40,* 291–298.

Long, B., & Konkle, T. (2017). A familiar-size Stroop effect in the absence of basic-level recognition. *Cognition*, *168*, 234–242.

Long, B., Konkle, T., Cohen, M., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, *145*, 95–109.

Long, B., Moher, M., Carey, S. E., & Konkle, T. (2019a). Animacy and object size are reflected in perceptual similarity computations by the preschool years. *Visual Cognition, 27,* 435–451.

Long, B., Moher, M., Carey, S. E., & Konkle, T. (2019b). Real-world size is automatically encoded in preschoolers' object representations. *Journal of Experimental Psychology: Human Perception and Performance*, *45*, 863–876.

Long, B., Yu, C. -P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*, E9015–E9024.

Makin, A. D. J. (2017). The gap between aesthetic science and aesthetic experience. *Journal of Consciousness Studies*, *24*, 184–213.

Merleau-Ponty, M (1962). Phenomenology of perception (C. Smith, Trans.). London: Routledge & Kegan Paul.

Orians, G. H., & Heerwagen, J. H. (1992). Evolved responses to landscapes. In J. H. Barkow and L. Cosmides (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 555–579). New York: Oxford University Press.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods, 51,* 195–203.

Ponce, C. R., Hartmann, T. S., Livingstone, M. S. (2017). End-stopping predicts curvature tuning along the ventral stream. *Journal of Neuroscience, 37,* 648–659.

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

Sensoy, Ö., Culham, J. C., & Schwarzer, G. (2020). Do infants show knowledge of the familiar size of everyday objects? *Journal of Experimental Child Psychology, 195*:104848, 1–13.

Srihasam, K., Vincent, J., & Livingstone, M. (2014). Novel domain formation reveals proto-architecture in inferotemporal cortex. *Nature Neuroscience, 17,* 1776–1783.

Van de Cruys, S., & Wagemans, J. (2011). Putting reward in art: A tentative prediction error account of visual art. *i-Perception, 2,* 1035–1062.

Wang., R., Janini, D., Kallmayer, A., Konkle, T. (2020). Mid-level feature differences support early EEG-decoding of animacy and object size distinctions. *Journal of Vision, 20,* 738.

Yonas, A., Pettersen, L., & Granrud, C. E. (1982). Infants' sensitivity to familiar size as information for distance. *Child Development, 53,* 1285–1290.

Yue, X., Robert, S., & Ungerleider, L. G. (2020). Curvature processing in human visual cortical areas. *NeuroImage, 222*:117295, 1–13.

Figure Captions

Figure 1. Example intact and texform images from Experiment 1, 2a, and 2b.

Figure 2. (a) Example displays of the size adjustment task in Experiment 1. (b) Results of the preferred visual size in pixels (y-axis), for big and small object images (x-axis), for intact and texform images (left and right panels). Each dot is from an individual subject, and the mean visual size for each condition is indicated with a horizontal line.

Figure 3. (a) Subjects completed 3 tasks in order in Experiment 2a and 2b: Size adjustment task on intact images, followed by an aesthetic 2-interval forced choice task, followed by a texform image recognition test. (b) Visual size preferences for the set of images whose texforms were not subsequently recognized, for Experiment 2a (left) and 2b (right). The y-axis shows the percentage of trials the subjects chose the canonical visual size (chance = 50%), plotted separately for intact and texform images. Error bars reflect 95% confidence intervals. (c) The plots are the same as in (b), but for the subset of items for which the texforms were subsequently recognized.

Appendix A: Texform Generation Procedure

Image collection

Images of objects and animals are collected from various sources including stimuli from previous works (Long, Yu, & Konkle, 2018; Konkle & Caramazza, 2013; Konkle & Oliva, 2012) and Google images. We first removed the background of the images, and then cropped it to the square bounding box of the objects or animals. Only images with resolution higher than 512 px × 512 px after this step were included, and repetitions of objects from the same basic level category were replaced. This results in a total of 180 images, with 45 big objects, 45 small objects, 45 big animals, and 45 small animals.

Normalization

All images were resized to 512 px × 512 px and converted to grayscale using the Rec. 601 Luma coding formula (red channel × 0.299 + green channel × 0.587 + blue channel × 0.114). They are then equalized across luminance and luminance histograms using the Spectrum, Histogram, and Intensity Normalization (SHINE) Toolbox (Willenbockel, Sadr, Fiset, Horne, Gosselin, & Tanaka, 2010). This step was done ignoring the background and without optimizing the structural similarity (SSIM) index (for details on the SSIM index, see Wang, Bovik, Sheikh, & Simoncelli, 2004). The images were then placed and centered on a gray (#828282) background of 640 px × 640 px.

Texform Generation

The texform images were generated with accelerated texform model (Deza, Chen, Long, & Konkle, 2019) modified from method used in previous studies (e.g., Long, Yu, & Konkle, 2018): The preprocessed intact images were placed in a simulated visual periphery as an input image to a metamer model (Freeman & Simoncelli, 2011). Then, a metamer image was

synthesized by iteratively coercing random noises to match the texture statistics of the input image for every overlapping simulated receptive field, in addition to roughly matching the structure given a low-pass residual of the input image. The procedure was run for 50 iterations using a variant of gradient descent, producing a final texform image, which is essentially a peripheral metamer of the intact image. The same intact image was passed into the model twice to generate a left and a right texform corresponding to the left and right visual periphery.

Appendix B: Texform Images Recognizability Pretest

To assess the recognizability of the texform images, we used a similar free-guessing method used in a previous study (Long et al., 2018). Thirty-six subjects from Amazon Mechanical Turk (Mturk) viewed the texform images and guessed what they were. (For a discussion of this pool's nature and reliability, see Crump, McDonnell, & Gureckis, 2013. All subjects were in the U.S., had an MTurk task approval rate of at least 95%, and had previously completed at least 50 MTurk tasks.) Half of the subjects were shown the 180 left texforms, and the other half seen the 180 right texforms. The texforms were shown one by one and the subjects simply typed their guesses in a textbox without time constraint. The subjects were instructed to always give one answer only (e.g., avoid answers like "A or B", "I don't know", or leaving the box blank).

Next, another 9 Mturk subjects (with the same Mturk qualifications) evaluated the guesses for the 360 texforms. Each subject viewed 40 original images (10 from each of the 4 types: big objects, small objects, big animals, and small animals) one by one with 18 guesses from 18 observers (20 images were paired with the guesses for the left texforms, and the other 20 were paired with the guesses for the right texforms) and judged whether each guess could be "used to correctly describe" the original image. They were asked to consider guesses correct as long as the guesses were descriptions of something of similar real-world sizes and shapes from the correct answers. The guesses were spell checked and corrected using Microsoft Word before showing to the subjects. We left the few guesses that were incorrectly and ambiguously spelled as is, since none of the likely interpretations of these guesses could influence the results. The count of guesses that were graded as correct yielded a recognizability score (ranging from 0 to 18) for each texform.

Appendix References

Crump, M., McDonnell, J., Gureckis, T. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8*:e57410.

Deza, A., Chen, Y. -C, Long, B., & Konkle, T. (2019, September 15). Accelerated texforms: Alternative methods for generating unrecognizable object images with preserved mid-level features. [Paper presentation]. *Conference on Cognitive Computational Neuroscience*, Berlin, Germany, https://ccneuro.org/2019/proceedings/0000879.pdf

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature neuroscience, 14*, 1195–1201.

Long, B., Yu, C. -P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences, 115,* E9015–E9024.