Published in final edited form as:
J Vis. ; 10(2): 18.1–1814. doi:10.1167/10.2.18.

Beauty and the beholder: Highly individual taste for abstract, but not real-world images

Edward A. Vessel and
Center for Neural Science, New York University, New York, NY, USA

Nava Rubin
Center for Neural Science, New York University, New York, NY, USA

Abstract

How individual are visual preferences? For real-world scenes, there is high agreement in observer's preference ratings. This could be driven by visual attributes of the images but also by non-visual associations, since those are common to most individuals. To investigate this, we developed a set of novel abstract, visually diverse images. At the individual observer level both abstract and real-world images yielded robust and consistent visual preferences, and yet abstract images yielded much lower across observer agreement in preferences than did real-world images. This suggests that visual preferences are typically driven by the semantic content of stimuli, and that shared semantic interpretations then lead to shared preferences. Further experiments showed that highly individual preferences can nevertheless emerge also for real-world scenes, in contexts which de-emphasize their semantic associations.

Keywords

visual preference; aesthetics; semantic association; individual differences

Introduction

Humans often get enjoyment from visual experiences, and can readily express preferences for some images over others. Whether in an art museum or driving down the road, some things we see reward us more than others. The study of visual preference and aesthetics, while extending back over centuries, has recently seen a renaissance using new experimental methods and based on a firmer understanding of the brain processes underlying visual processing and emotional responses.

Can preferences for visual stimuli be predicted? Researchers of visual aesthetic responses have placed varying emphasis on the importance of stimulus features versus internal factors. The majority of empirical studies of preference have sought to identify the common factors determining preference by testing whether average image preference for a group of observers can be reliably predicted from various factors in the stimulus set (Aitken,1974; Bar & Neta, 2006; Berlyne, 1971; Berlyne & Ogilvie, 1974; Eysenck, 1940; Martindale, Moore, & Borkum, 1990). Investigations of preference from an evolutionary perspective have focused on such common determinants of preference, stressing high agreement across

Corresponding author: Dr. Edward A. Vessel., ed.vessel@nyu.edu. Address: 2-4 Washington Pl., Rm 156B, New York, NY 10003, USA..

Commercial relationships: none.

individuals (Kaplan, 1992; Kaplan & Kaplan,1995; Orians & Heerwagen, 1992). Indeed, studies from these schools of thought have found a number of stimulus properties that are powerful determinants of preference for most observers. Berlyne's influential psychobiological approach emphasized the role of objectively measurable stimulus variables and "collative" properties (e.g. novelty, complexity, surprise; (Berlyne, 1958, 1970). Facial attractiveness is one obvious—albeit highly domain specific-example of high commonality across observers, with reports of cross-observer agreement as high as 0.5 thought to be driven by facial symmetry, similarity to an average face, and the presence of secondary sex characteristics (Bronstad & Russell, 2007; Gangestad, Thornhill, & Yeo,1994; Rhodes, Sumich, & Byatt, 1999; Thornhill & Gangestad, 1999). Turning back to a more general domain, the landscape assessment literature contains numerous studies in which observers tend to like the same scenes, with agreement values ranging from 0.4 to 0.6, with a large degree of this shared variance in preference attributed to stimulus factors such as naturalness, complexity, vista, mystery, coherence, legibility and refuge (up to 80% of the variance in mean preference ratings; Kaplan & Kaplan, 1995). A recent study of a wider variety of real-world scenes also showed high agreement across observers (Yue, Vessel, & Biederman,2007). The explanatory power of such factors may stem from the survival utility of a particular environment or viewpoint—it has been argued that biases toward developing preferences for certain environmental traits (e.g. the presence of resources or the ability to see oncoming predators) may have become genetically encoded over the course of evolution and therefore generally shared amongst all humans (Appleton, 1988; Kaplan, 1992; Orians & Heerwagen, 1992; Wilson, 1993). Finally, a host of additional stimulus features that appear to have at least modest effects on group preferences have been identified using this approach, such as contour shape (sharp vs. curved; Bar & Neta, 2006), contrast and clarity (Reber, Schwarz, & Winkielman, 2004; Tinio & Leder,2009), color (McManus, Jones, & Cottrell, 1981), fractal dimension (Aks & Sprott, 1996; Graham & Field, 2007; Van Tonder, Lyons, & Ejima, 2002), number of sides (e.g. "complexity"; Aitken, 1974), aspect ratio (McManus,1980), symmetry (Rentschler, Jüttner, Unzicker, & Landis, 1999) and stimulus prototypicality (Shortess, Clarke, Richter, & Seay, 2000). Thus, the picture that emerged from these studies is that measuring group preferences for stimuli allows one to predict the preferences of new observers.

The above findings would appear to call into question the old adage that "beauty is in the eye of the beholder"—yet perhaps such a conclusion is premature. The tendency to average results across observers, stressing group norms over the expression of individual taste, de-emphasizes the potential role that factors internal to the observer may have on preference. An alternative approach posits that the relation between preferences and stimulus attributes, when present, may be indirectly mediated via latent variables. According to this hypothesis, the proximal determinants of preferences are internal factors reflecting subjective, and therefore inherently individual, evaluations. This approach does not deny the existence of general principles underlying preference formation—however, such principles are hypothesized to exist at an internal process level, leaving room for individual differences. Any common preferences observed across individuals would be the result of common experiences with a stimulus, not stimulus attributes *per se*.

One prominent theory of this type is Martindale's cognitive theory of aesthetic preference, which is based on neural network concepts and stresses the importance that meaning plays in determining preference (Martindale, 1984). Stimuli that activate increasing numbers of cognitive units in a hierarchically constructed network lead to greater preference. A related hypothesis is that preference is a result of perceptual fluency (Reber et al., 2004). In this framework, stimuli that are processed easily are more preferred, and ease of processing depends on an observer's individual history. Biederman and Vessel (2006) postulate that preference formation is a result of the interplay between subjective novelty, e.g. how new a

stimulus feels to an observer, and how well the observer is able to make sense of a stimulus and relate it to previous knowledge, which they term interpretability. For example, a specific visual noise pattern may be objectively novel, but not feel new to an observer (and not be particularly liked) due to the lack of interpretable elements. A scene of familiar objects presented in a new arrangement, however, contains both an element of subjective novelty and is relatable to previous knowledge, potentially leading to a preference. In Biederman & Vessel's theory, the effects of these factors on preference are realized through the activation of later stages of perceptual/associative processing, with novel, highly interpretable stimuli leading to greatest activation of these regions. Similarly, Silvia casts preference, interest, and other 'knowledge-based' emotions in terms of appraisal theory, again identifying a novelty-related component (appraisal of novelty) and a meaning-related component (appraisal of coping potential; (Silvia, 2005a, 2005b).

An emphasis on the internal aspects of preference formation requires the use of methods that allow the experimenter to relate individuals' preferences to putative subjective, internal factors. And yet, many of the tests of the theories above have relied on designs requiring the averaging of data over observers (e.g. Berlyne, 1970; Kaplan & Kaplan, 1995; Martindale et al., 1990; Reber, Winkielman, & Schwarz, 1998; but see Discussion for exceptions). In this paper, we focus on measuring agreement across observers, which allows for an assessment of the relative contributions to preference of individual versus common factors.

When can we reasonably expect that people will agree in their preferences, and when must we sacrifice the statistical power that averaging provides and focus on individual preferences instead? We set out to measure the extent to which semantic information plays a role in determining whether people will agree in what they like and dislike. A number of studies have explicitly manipulated (or measured) the "meaningfulness" of stimuli to investigate its relationship to preference, with mixed results (Cupchik & Gebotys, 1990; Martindale et al.,1990; Millis, 2001; Russell, 2003). However, the relationship between meaning and *agreement* has not been tested directly. We hypothesized that if preference formation is indeed based on internal factors, then shared preferences may emerge as a consequence of shared semantic interpretations of stimuli.

To assess across-observer agreement in visual preferences and the possible role of the semantics of scenes in determining those preferences, we had observers make pairwise preference judgments on pictures of real-world scenes and of abstract images such as fractals and kaleidoscopic patterns (Figures 1a and 1b respectively). Importantly, we utilized large, heterogeneous image sets, which was necessary to insure that individual differences in preference, if present, could emerge. The use of such large image sets required the development of a novel, adaptive paired-comparison method to robustly measure preferences across a population of observers. In Experiment 1, preferences were measured for real-world scenes and abstract images in separate sessions. In Experiment 2, the two stimulus types were intermixed to allow direct comparison.

If stimulus variables are the dominant determinant of preference, there should be high agreement across observers for all stimuli—abstract as well as real-world scenes. Adding meaning would not have a systematic effect on agreement, although it may change the mean level of preference (either increasing or decreasing it; Berlyne, 1970; Bornstein, 1989; Zajonc, 1968).

On the other hand, if the proximal determinants of preferences are internal, then one would hypothesize that real-world scenes should show higher agreement across individuals than abstract images. This is because the shared semantic associations would inevitably lead to more similar internal states in response to the real-world stimuli. For example, in the real-

world images in Figure 1a, high agreement between observers may emerge due to the compelling association of the left image with "leisure" and the right image with "work," whereas analogous associations are not available for the abstract images in Figure 1b.

Note that the hypothesis above is very different from asking what the effect of adding semantics might be on the mean level of preference for the two stimulus types. A possibility to consider is that meaning in and of itself may contribute to the formation of positive preference, as predicted by Martindale's theory and perceptual fluency. If this were the case, then virtually all real-world scenes should be preferred over abstract images, leading to a large difference in the mean preference for the two groups. This hypothesis does not predict any difference in across-observer agreement for the two stimulus types.

On the other hand, the effect of adding meaning on the mean level of preference may not be so easily predictable. If meaning is more heavily weighted than stimulus factors in determining preference but does not alone predict the direction of this weighting (as predicted by Biederman & Vessel's hypothesis and consistent with the "internal" factor hypothesis above), the effect of adding meaning will depend on the exact images in the stimulus set. Particular real-world images may be selected which are universally liked or disliked, such as photographs of baby animals or mutilated bodies, respectively.

In Experiment 1, we found that abstract images showed significantly lower agreement across observers than did real-world scenes. In Experiment 2, we show that the mean levels of preference for our abstract and real-world stimulus sets overlap to a very large degree, and importantly, show no difference in their range of preferences. The results of these experiments reveal that shared semantic knowledge leads to shared preferences: people agree in what they like and dislike when the stimuli and decision context lead to similar assessments of a stimulus' *meaning*. Our results rule out theoretical approaches which seek to base preferences solely on image features. While images may contain certain features that activate a core of common experiences shared by all observers, they also engender unique associations for every individual, leading to divergent tastes.

Methods

Stimuli

The abstract images consisted of 96 images in six sets of 16 each (see Supplementary Figure S1 for the entire set of abstract images): (i) abstract shapes created using Maya 3D rendering software (http://www.alias.com); (ii) kaleidoscopic images constructed by reflecting a sliver of a real-world image about a number of symmetry axes; (iii) pseudo-colored electron microscope images obtained from the Centre for Microscopy and Micro-analysis at The University of Queensland (http://www.uq.edu.au/nanoworld); (iv) fractal images created with publicly available interactive programs (http://sprott.physics.wisc.edu/fractals.htm, http://www.fractint.org, http://www.arosmagic.com/Fractals); (v) satellite imagery courtesy of the U.S. Geological Survey (http://edc.usgs.gov and http://landsat.usgs.gov); (vi) an "other" category of images collected from public Internet sources. The 96 real-world scenes were selected from a set of images used in Vessel and Biederman (2002) to vary on a large number of dimensions, such as natural vs. urban, depth (close-up vs. vista), and the presence of objects, people, and animals (see Supplementary Figure S2 for the entire set of real-world scenes). All images were 300 × 300 pixels and were specified in RGB color space. The images were shown on a Viewsonic P95f+ CRT display controlled by an Intel-based PC running Matlab 6.5 (MathWorks Ltd., Natick MA) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). The stimuli were presented on a gray background (7.0 cd/m^2) and subtended 9.6° visual angle.

Observers

All observers had normal or corrected vision and were naïve to the purpose of the experiment and to the stimuli. They consented to participate in accordance with NYU human subjects policies, and were paid for their participation.

35 observers participated in this study. Experiment 1 included 11 observers (7 female, 2 left-handed) with a mean age of 28.5 ± 7.5 years. Experiment 2 included 13 observers for Set 1 (7 female, 1 left-handed), with a mean age of 26.1 ± 6.6 years, and 11 observers for Set 2 (7 female, 1 left-handed) with a mean age of 23.0 ± 4.3 years. Five additional observers were excluded fromExperiment 2 because they produced low (below 0.5) within-observer reliability (1 for Set 1, 4 for Set 2).

Single draw paired comparisons

This method is a modification of a standard paired comparison method. Images were shown one at a time and observers were requested to compare each new image to the immediately preceding image.

On each trial, observers saw a fixation point for 300 ms followed by an image for 1 s. After the image disappeared, they indicated whether they preferred the image just seen or that shown on the preceding trial by pressing either the right (prefer current) or left (prefer previous) arrow key. This required that observers remember the image shown on each trial for comparison on the next trial, which observers readily did after a few practice trials. The observer's response on each trial was used as input to a sorting algorithm, heapsort (Williams, 1964), which adaptively chose the next image for presentation to minimize the total number of comparisons required (see Appendix A for details). A block of trials ended when the algorithm determined that a first pass sorting had been achieved (278 ± 20 trials, on average). Each observer was run in a total of 4 blocks, either in succession or over a period of one day.

This adaptive algorithm was used as an alternative to performing a full set of comparisons between every pair of images. The difference between the two approaches can be understood intuitively with examples from how sports tournaments can be organized. A full comparison is analogous to a "round robin" tournament where each team plays every other team, and the heapsort algorithm is analogous to a "playoff" tournament.

The heapsort algorithm was used to guide trial selection, but not for inferring the preference order. Given that observers make noisy judgments (there is no guarantee that their preference judgments are transitive, or even that they produce the same response upon repeated presentations), the set of comparisons was used to *estimate* preferences rather than to sort. The data from the 4 blocks were used to create comparison matrices for each individual block as well as one matrix for the entire experiment. The observers' underlying stimulus preferences were estimated from these matrices for the entire experiment, for each individual block, and for the first and second half of the experiment. Using logistic regression, we computed a vector of image preferences most consistent with the experimentally observed choices (see Appendix A for details).

The validity of the single draw paired comparison procedure was verified by comparing those results with preference ratings obtained for each stimulus type separately (for a subset of subjects from Experiment 1), and by visually examining the full set of choices that observers made in the first experiment for consistency (see Figure 3).

The instructions to the observers stressed that there were "no right or wrong" answers, only their "subjective impressions of which pictures they liked better than others." They were

told that a good way to think about each comparison was to consider which picture they would rather see again, or which one they would rather choose as a poster to put in their room. Observers were told to base their responses on their feeling about the image *at that moment*, regardless of choices made on previous trials. Before the start of the experiment, observers were given a short practice of 12–20 trials to familiarize them with the single draw method and the types of images they would be seeing. Throughout each block, observers were given a short break after every 60 trials.

Results

In the first experiment, observers made preference judgments for 96 abstract images in one session and 96 real-world scenes in a separate session, collected on a later day. We used a modified paired comparison method in which images are shown one at a time and each is compared to the one immediately preceding it ("single draw"; see Methods above). The across-observer agreement was quantified by computing the pairwise correlation between the preference estimates of every pair of observers. The within-observer reliability was quantified by computing the correlations between the preference estimates derived from data of the first and second half of each observer's session.

Experiment 1

The across-observer agreement was much higher for real-world scenes than for abstract images ($N = 55$, $ks = 0.67$, $p < 10^{-11}$; Figure 2). The pairwise correlations between different observers' preferences for the real-world images had a mean of 0.46 (±0.02 *SE*), whereas the correlations between preferences for abstract images averaged 0.20 (±0.02 SE). Importantly, the low agreement for the latter set was not a result of poor reliability of individual observers' preferences for abstract images: the within-observer correlations show that individuals were as consistent in their preference judgments of abstract images as they were for real-world scenes (means 0.67 and 0.71 for the former and latter image sets, respectively; $N = 11$, $ks = 0.45$, $p = 0.15$). The matrices of choice probabilities (Figure 3) provide a means to visualize more details of the results, again revealing high within-observer reliability for both stimulus types (central panels, Figures 3b and 3e) alongside across-observer agreement that is high for real-world scenes (bottom-right, Figure 3f) but low for abstract images (top-right, Figure 3c). Similar results were obtained in a rating experiment, where the across-observer correlation was 0.41 for real-world scenes and 0.20 for abstract images ($N = 28$, $ks = 0.54$, $p < 0.0004$; data not shown). The rating experiment again showed high within-observer agreement for both image types, even higher than that obtained with paired comparisons (0.86 and 0.85 for real-world scenes and abstract images respectively; this is likely due to the preferences being measured directly in the ratings experiment, compared with having to estimate them from half of the data set in the single draw method. At the same time, rating methods are limited in other important aspects; see below and Experiment 2.)

The marked difference in levels of agreement for the two sets of images suggests that observers' visual preferences were driven by the semantic content of images, when it existed. Since the semantic interpretations evoked by real-world scenes are shared by different individuals, they lead to shared likes or dislikes for those images. In contrast, abstract images do not evoke commonly shared semantic knowledge, resulting in marked reductions in agreement between observers' judgments about the visual appeal of those images.

Until now there was no direct evidence that the visual preferences are indeed mediated by such semantic associations. Instead, one might have hypothesized that there are systematic

differences in the colors, shapes or textures of scenes related to “good” and “bad” semantic knowledge, and that the high agreement in preferences across observers is driven directly by those putative visual differences, rather than being mediated indirectly by the semantic associations the images invoke. This alternative hypothesis is ruled out by our finding that the agreement between observers drops dramatically when they form preferences about abstract images, since those images also contained a wide range of colors, shapes and textures, yet did not lead to high preference agreement. The purpose of including abstract images from six different categories was to span as wide a range of preference as possible with the set of 96 images. An analysis by category revealed no consistent effect of abstract image category on across-observer agreement. However, it is possible that the small number of images in each category limited our ability to observe such an effect.

A remaining concern is that the range of preferences spanned by abstract images may have been significantly narrower than that spanned by real-world scenes. In sensation and perception, subjects’ ability to compare stimuli along a given dimension decreases as the distance between stimuli is reduced along this dimension (Fechner, 1912; Weber, 1846). If the underlying visual preferences for abstract images spanned a narrower range than that spanned by real-world scenes (on a common, absolute scale), this could have similarly reduced subjects’ ability to compare those preferences reliably. For this to happen without an accompanying drop in the within-subject reliability, subjects would have had to rely on memory of their responses to maintain consistency; this may be possible given the known high capabilities of human memory for complex pictures (Shepard, 1967; Standing, 1973).

At first glance, it may seem that obtaining ratings of visual preferences for both types of stimuli on a common scale would address this issue. However, rating methods have severe limitations in this regard, because they do not require observers to make direct comparisons between stimuli of different types on single trials, thus allowing them to maintain multiple, non-interacting scales for different types of stimuli. It has been shown that observers indeed have a strong tendency to maintain independent mappings to category responses (such as rating scales) even when two classes of stimuli are intermixed trial-by-trial (Parducci, Knobel, & Thomas, 1976). This would be particularly troublesome when the goal is to compare between responses to different types of stimuli. The possibility that observers could maintain separate mappings to separate rating scales for abstract images and real-world scenes rules out the use of a rating scale for comparing across such different categories of images. (Additional problems with rating methods include possible criterion changes over time, floor and ceiling effects, and effects of non-normal stimulus distribution on response category assignment) (Adolphs & Tranel, 1999; Parducci & Wedell, 1986). In order to obtain reliable data about relative preferences, we therefore intermixed our two highly different classes of stimuli and used the single-draw paired comparison procedure to obtain preference judgments about them.

Experiment 2

In the second experiment, observers made direct preference comparisons between abstract and real-world images. We used sets of 96 images comprised of 48 abstract and 48 real-world scenes presented in an intermixed randomized order (i.e., they made direct comparisons between an abstract image and a real-world scene on half of the trials, on average). Pilot experiments indicated that intermixing may affect the level of agreement in preference judgments of the real-world scenes (compared to when presented alone; see also below). Therefore, the second experiment was run on two separate groups of observers who were given the same 48 abstract images but different sets of 48 real-world scenes. Specifically, the 96 real-world scenes from the first experiment were split into “high-” and “low-agreement” groups (Sets 1 and 2, respectively).

The abstract and real-world images were found to span comparable ranges of preferences for both experimental sets (Set 1: ranges 9.6 and 10.3 for abstract and real-world images, respectively; paired $t[12] = 1.2$, $p = 0.25$, 13 observers; Set 2: 9.75 vs. 10.3; paired $t[10] = 0.86$, $p = 0.43$, 11 observers). These results rule out the potential concern that the low across-observer agreement in preference for abstract images was due to those images spanning a narrower range of preferences than that spanned by real-world scenes. At the same time, the results did reveal a difference in the mean preferences for the two types of stimuli, with the real-world scenes, on average, being preferred to the abstract images (Set 1: paired $t[12] = 5.44$, $p < 10^{-3}$; Set 2: paired $t[10] = 3.08$, $p = 0.012$). Further research is needed to test whether these differences arise from familiarity, semantic associations, the specific choices of stimuli we used, or a combination of these factors.

A surprising consequence of the intermixing in Experiment 2 was a dramatic drop in across-observer agreement for the real-world scenes compared to Experiment 1. Agreement of preference judgments for the subset of real-world images in Set 1, which consisted of images yielding high agreement in the first experiment (mean pairwise correlation 0.60), decreased to a correlation of 0.27 when intermixed (Figure 4, central panel; $N1 = 55$, $N2 = 78$, $ks = 0.57$, $p < 10^{-9}$). Set 2, which consisted of images that yielded relatively lower agreement in the first experiment, also showed a significant reduction from 0.29 to 0.18 (Figure 4, central panel; $N1 = 55$, $N2 = 55$, $ks = 0.28$, $p = 0.027$). The differences in the pairwise correlations of subjects' preferences are clearly seen in both the means and the distributions from the two experiments (Figure 4, center and side panels respectively). The decreased across-observer agreement in the intermixed (second) experiment was not a result of reduced within-observer consistency, which was high for both sets (0.65 for Set 1, and 0.70 for Set 2).

Discussion

We obtained visual preference judgments for two stimulus sets—real-world scenes and abstract images—and measured for each the between-observer agreement as well as the within-observer reliability. Confirming previous studies, there was a high degree of agreement in observers' visual preferences for real-world scenes. In contrast, the across-observer agreement was markedly lower for abstract images, while at the same time the within-observer reliability for this stimulus set was comparable. These results indicate that common preferences observed for real-world scenes are a result of observers' shared semantic interpretations. When no clear semantic content is present in the images, visual preferences emerge as highly individual, yet no less robust as those for real-world scenes. Finally, in a second experiment we confirmed that the range of preference for abstract and real-world images was comparable by obtaining preference data when the two types of stimuli were compared directly.

An alternative interpretation of the results of Experiment 1 one may propose is that the main difference between the real-world scenes and the abstract images is that the former contain more information, and that additional information of any nature would lead to an increase in agreement. To test this possibility one would need to augment the abstract images with additional information that would be comparably perceptually salient to object identity, but not contain semantic associations. For example, one could test whether consistently pairing abstract images with arbitrarily chosen sounds or smells would lead to an increase in agreement. Note that in order to test this in a controlled way, one would need a method for quantifying the extra "bits of information" present in real-world scenes and in the added sounds/smells in a formal way, and also to evaluate the potential "congruency" of any added information paired with an abstract image. While it may be possible to meet these experimental challenges, the interpretation we proposed for the data presented here make a

clear prediction: if the new information were added in such a way as to *not* add semantic information, then agreement would not increase.

Several studies within the field of experimental aesthetics have highlighted the critical importance of "representational" versus "abstract" styles for the perception of paintings using a multi-dimensional scaling approach (Berlyne & Ogilvie, 1974; Cupchik, 1974). Our results, which were obtained using a non-artistic stimulus set vastly larger than any previous study, show that the major effect of increasing semantic associations is not to universally increase preference, but instead to increase the degree to which different observers agree in which images are liked or disliked.

An unexpected finding of Experiment 2 was that, when real-world images are intermixed such that they are compared directly with abstract images, between-observer agreement about visual preference of those images is significantly decreased (cf. Figure 4). What could be the cause of this dramatic effect of intermixing the two types of stimuli? A reasonable scenario is that during the real-world scene session of Experiment 1, observers could allow associated meanings to inform their preference choices (either intentionally or unconsciously), since all images contained such meaning. In contrast, in Experiment 2 two different image categories were intermixed, which forced observers to make direct comparisons between an abstract image and a real-world scene on approximately half of the trials. In this situation, relying on meaning to make preference choices would make the task more difficult, since this dimension was not present for many of the stimuli. Therefore, a reasonable strategy would be for the observers to de-emphasize this dimension of meaning —a strategy which affected even the one-quarter of comparisons where both stimuli had readily available semantic interpretations. Thus, we propose that in the 'intermixed' condition observers made much less use of semantic information when making their preference judgments in all trials (and not just the trials that contained abstract images), and instead based their decisions more on the visual aspects of the images. It may be possible to provide more direct evidence for this interpretation by testing how systematic variations in the ratio of abstract to real-world images, and/or the ratio of across-category versus within-category comparisons, affect agreement between observers' preferences on the subset of real-world images.

Indeed, context and attention have previously been shown to mediate the influence of other factors on preference judgments. For example, the extent to which an event's duration influences its reported preference can be affected by whether the task draws attention to duration or not (Ariely, Kahneman, & Loewenstein, 2000; Ariely & Loewenstein, 2000; Fredrickson & Kahneman, 1993). Interestingly, the effect of context in our experiment carried over even to trials during which observers had to choose the more preferred between two real-world scenes (reanalysis not shown). There was also a marginally significant decrease in agreement for abstract images when intermixed, possibly reflecting the decreased influence of factors that promoted across-observer commonality for abstract images (e.g. symmetry, color) when judged alone.

The single draw paired comparison method we developed for these studies represents an important addition to the toolbox of methods available for studying preference. Paired comparisons allow for stimuli of very different kinds to be directly compared, whereas when using rating scale methods there is a concern that observers may maintain separate mappings to separate rating scales, as has been shown to occur (Parducci et al., 1976). Furthermore, the paired comparison method does not pre-suppose that responses to complex and diverse stimuli can be neatly mapped onto a single scalar dimension. For example, data obtained with this method could, in principle, reveal violations of transitivity, where participants systematically place A above B and B above C, but also place C above A. Such perceptual

or cognitive effects could not, by definition, reveal themselves when using a rating method. In our data set, observers' paired comparisons responses could be explained reasonably well as arising from a single scalar dimension of preferences, and it is worth noting that, thanks to our use of a paired comparison method, this is an observation emerging from our data, rather than a circular outcome of the choice of method. At the same time, there are several noteworthy disadvantages. Studying whether and how preferences change with repeated presentations and/or over time may be difficult to do with paired comparisons, and probably better done using a rating scale design. At a practical level, a "full" paired comparisons design requires many more trials than direct ratings ($O[N^2]$ compared with $O[N]$ respectively). Careful selection of a sorting algorithm to guide trial selection can alleviate this problem some-what—e.g., our use of the heapsort algorithm drastically reduced the number of trials needed to achieve reliable preference estimation—but the number of trials required is still going to be higher (see Appendix A). A related complication is that, even when it is appropriate to do so, collapsing the paired comparison responses to ordinal lists, either along a scalar or a multi-dimensional space, requires the use of estimation procedures (such as the logistic regression we have used here), and any such analysis unavoidably adds a layer of assumptions to the raw data. For these reasons, it is probably advisable to stick with rating scale procedures when concerns like those mentioned above are not central, and reserve paired comparison designs to cases that involve comparisons of "apples with oranges", and/or for complex multidimensional stimuli when a scalar ordinal scale should not be presupposed.

The strikingly low agreement for abstract images found in our experiments argues strongly against the idea that objective visual features play a major role in directly determining preference, as hypothesized by Berlyne (1971). Instead, our findings provide support to theories that postulate that internal factors drive preference and that when stimulus features play a role it is via the mediating role of meaning (Biederman & Vessel, 2006; Leder, Belke, Oeberst, & Augustin, 2004; Martindale, 1984; Silvia, 2005a, 2005b). More generally, our findings underscore the importance of explicitly measuring individual differences in preference, an approach that has been successfully applied in the context of attractiveness (Hönekopp, 2006), simple figures and colors (McManus, 1980; McManus et al., 1981). Specific predictors of individual differences have also been investigated by a number of groups (Eysenck, 1995; Rawlings, 2000; Yu & Shepard, 1998; Zuckerman, Ulrich, & McLaughlin, 1993).

Although our findings suggest that the presence of semantic information is heavily weighted in determining preferences, it should be noted that preference is not a monotonic function of "meaningfulness"—a finding which is contrary to Martindale's hypothesis (Martindale, 1984). Rather, the semantics of an image can induce either high *or* low preference. In Experiment 1, the high agreement was driven by low preference images as much as by high preference images, and in Experiment 2, there were real-world scenes that were overall less preferred than many of the abstract images. These findings are a better fit with the neurocomputational hypothesis of Biederman and Vessel (2006), which suggests a strong influence of associative activity.

Major theories of emotion recognize the central role that evaluation of an event's pleasantness play in determining one's emotional response (Ellsworth & Scherer, 2003). Our results are in line with a view of preference as a "knowledge-based" emotion such as interest and curiosity, and may be understood as resulting from a series of appraisals of an event (Silvia, 2005a, 2005b). But, in the narrower context of our forced-choice paradigm, preference could also be thought of as the process of collapsing a multi-dimensional emotional space onto a scalar axis, to allow choice between the widely different options the environment offers for action, attention and consumption. Our stimulus set was purposefully

selected to include a fairly narrow range of emotional reactions and did not include strongly arousing images such as gore or sexual content. Thus, although our results show consistency and a robust range of preferences, the experiments reported here highlight the fact that the same stimulus can lead to different emotional reactions in different observers, which may well be caused by individual differences in the appraised meaning, novelty, or coping potential of a perceptual event. It should also be noted, though, that our experimental stimuli were not designed (and are not well suited) to explore the relationships between preference and emotion in general.

A consequence of the high agreement in preferences for real-world images is that it allows one to reliably predict which of those images will be highly preferred by new observers, and which will fare poorly (see also Biederman & Vessel, 2006; Eysenck, 1940; Kaplan & Kaplan, 1995; Martindale, 1984; Vessel & Biederman, 2002; Yue et al.,2007; note that such predictions will be much less reliable for abstract images). This conclusion is supported when we calculate Cronbach's alpha, a measure of test reliability often reported in the aesthetics literature—for real-world scenes in Experiment 1, $\alpha = 0.90$, while for abstract images $\alpha = 0.72$. For illustration of such "high" and "low" preferred images, the scenes in Figure 1a are arranged by the order of preferences obtained from our observer set (left and right panels, respectively). Examination of these images offers an intuitive sense of how semantics may affect visual preferences. For example, the associations of calm and relaxation evoked by the Japanese garden may play a role in the higher preferences obtained for this image compared with the parking lot, which is likely to evoke associations of work and/or errands.

In addition to associations formed at a conceptual level, previous studies suggest that associations determining a shared preference can be formed on the basis of perceptual, or even sensory experiences, such as interactions with smooth versus sharp objects (Bar & Neta, 2006). Exposure to social and cultural values about what is attractive may also be a source of association leading to shared preferences (Bronstad & Russell, 2007). The influence of strongly diagnostic low level visual cues (such as sharpness in the case of objects, or secondary sex characteristics in the case of faces) may explain why, even for abstract images, the average correlation between observers does not drop to zero.

Conclusions

Taken together, our results show that the high degree of agreement observed for real-world scenes is a result of observers' shared semantic interpretations, but also that the effect of shared semantic experience can be greatly reduced by context. Further research is needed to determine whether the motor and visceral reactions normally associated with preferences (e.g., visual orienting, facial muscle movement, feelings of pleasure) would also show such sensitivity to context (Shimojo, Simion, Shimojo, & Scheier, 2003; Winkielman & Cacioppo, 2001). The unveiling of highly individual preferences in non-semantic contexts may have implications in a broader social sense: despite continuous exposure to increasingly global trends in fashion, design, and visual media, people reveal robust, individual tastes when tested in the right context.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Eye Institute grants EY07158-03 (EV) and EY14030 (NR). We thank M. Goldberg, R. Shapley, A. Shpiro, and M. Fukui for comments on earlier versions of this manuscript. We thank J.

Lobo for suggesting the heapsort algorithm. We are grateful to J. Brenneman and J. Silver for help with stimulus design, C. Sprott for the use of his fractal database, and D. Waddell for supplying the set of electron microscope images.

Appendix A

Adaptive trial selection using heapsort

Image presentation for the single draw paired comparisons method was guided using an adaptive algorithm based on heapsort (Williams, 1964). The process of assessing image preferences by paired comparisons is conceptually very similar to a sorting procedure. The theory of computing tells us that sorting should take $O(N\log N)$ trials, which, for a large number of images, is significantly less than the $O(N^2)$ trials needed to compare every image against every other image (for N = 96, this translates to 632 vs. 4560 comparisons).

We sought a trial selection algorithm that would lead to a relatively flat distribution of presentations across all stimuli and was compatible with our single draw method of presenting only one image at a time. Many different sorting algorithms exist which differ in their detail (although they all take $O(N\log N)$ steps). When visualized as a binary tree, sorting by heapsort allows one to focus on local comparisons within a binary tree structure without continually revisiting the same nodes.

The heapsort algorithm functions in two stages. In the first stage, a string of n items to be sorted are first arranged into a structure called a *heap*. A heap is a binary tree, with the root at the top, in which the value of any node (e.g. where two branches meet) is greater than any of the leaves below it (e.g. the entire subtree which emanates from that node). A heap is created by placing items along the bottom of a tree, and then as items are placed in the next row up, pairwise comparisons are made between each of the leaves on the bottom level and a swap is made if the value of the leaf is higher. This process continues as the remaining levels of the tree are created, so that by the time the root node is reached (in at most $n\log_2 n$ steps), the highest valued item is at the root.

The second stage of the heapsort algorithm takes this heap as its input, iteratively removing the root item and restoring the heap property of the remaining nodes until a full sort has been achieved, in at most $2n\log_2 n$ steps, or $O(n\log_2 n)$.

Since the values we are sorting are subject to noise (e.g. our observers may not always be consistent in their choices) and we are not guaranteed that preferences are transitive, achieving a complete sort is impractical. Therefore, we do not use heapsort to perform a full sort, but instead only use the first half of the algorithm to guide the formation of a heap. The choices a subject makes on each trial are passed to the matlab function sort_heap_external (Nijenhuis & Wilf, 1978) for selection of the next comparison to be made, and are also stored in a comparison matrix **M**. In addition, the heapsort algorithm does occasionally call for a comparison between two images, neither of which were shown on the previous trial. In this case, an additional comparison is inserted which is not passed to the sort_heap_external function, but is still stored in **M**.

One block of trials ended when the algorithm signaled that the items were sorted into a heap. Each subject was run in four blocks of trials. Two of these used an initial sorting order which was a mirror reversal of the other two. Two blocks used the heapsort algorithm sorting normally (from lowest to highest), while two blocks operated as if sorting from highest to lowest, crossed with initial sorting order. These manipulations produced a more evenly distributed presentation of images. The most presented image was shown, on average, 2.0 times more than the least presented image. However, since different observers

did not always prefer the same images, the overall stimulus presentation showed a flatter distribution, where the most presented image was shown an average of 1.5 times the least presented image.

Estimation of preferences using logistic regression

Once a large number of comparisons have been made, the resulting comparison matrix (**M**) is used to estimate the underlying scale values of the stimuli. We performed this estimation using a logistic regression procedure implemented from McGuire and Davison (1991).

This approach is based upon the BTL choice model (Bradley & Terry, 1952; Luce, 1959), which states that the proportion (p) of times one stimulus (k) is chosen over another stimulus (j) is related to their scale values x_j and x_k by:

$$p_{jk} = \frac{e^{(x_k - x_j)}}{1 + e^{(x_k - x_j)}} \qquad 1 \leq j, \quad k \leq J. \quad \text{(A1)}$$

A regression is set up as a series of equations in which the vector of scale values x is multiplied by a matrix of dummy variables d, summed, and set equal to each element in L, the logit of each element in **M**:

$$L_{jk} = \sum_{i \neq 1} x_i d_i \qquad i = 2, \ldots, J, \quad \text{(A2)}$$

where

$$L_{jk} = \ln\left(\frac{p_{jk}}{1 - p_{jk}}\right) \quad \text{(A3)}$$

This logit transformation linearizes the regression. The simultaneous minimization of this set of linear equations (using the Matlab function lsqnonlin) produces a vector of scale values x which maximizes the likelihood of the observed comparison matrix.

There are two modifications that were required to make this algorithm useful for our purpose:

1. When using an incomplete comparison matrix as input, there will be entries for which both **M**(j, k) = 0 and **M**(k, j) = 0. This cell therefore contributes no data towards the estimation, and causes an error in the computation of L. As a solution, this equation is dropped from the estimation.
2. When there is unanimous agreement for a comparison, **M**(j, k) = 1 and **M**(k, j) = 0. The logit L will be $\pm\infty$. As an alternative, a number very close to 0 or 1 is substituted for **M**(j, k) and **M**(k, j), resulting in a very large or very small (but finite) logit.

One problem with this solution is that solving the set of linear equations requires that the mean stimulus scale value be set to zero to avoid a singular design matrix. The resulting values carry no information about absolute scale. Therefore, it is impossible to make absolute comparisons between the resulting estimated preferences across separate experiments. However, comparisons of ordinal or correlational statistics across experiments are valid, as are comparisons of estimated preferences *within* an experiment.

Distribution testing

Normal probability plots of the distributions of across- and within-observer correlations indicated deviations from normality. Therefore, Kolmogorov-Smirnov tests were used to assess the similarity of two distributions rather than t tests.

References

Adolphs R, Tranel D. Preferences for visual stimuli following amygdala damage. Journal of Cognitive Neuroscience. 1999; 11:610–616. PubMed. [PubMed: 10601742]

Aitken PP. Judgements of pleasingness and interestingness as functions of visual complexity. Journal of Experimental Psychology. 1974; 103:240–244.

Aks DJ, Sprott JC. Quantifying aesthetic preference for chaotic patterns. Empirical Studies of the Arts. 1996; 14:1–16.

Appleton, J. Prospects and refuges revisited. In: Nasar, JL., editor. Environmental aesthetics. Cambridge University Press; New York: 1988. p. 27-44.

Ariely D, Kahneman D, Loewenstein G. Joint comment on “when does duration matter in judgment and decision making?” (Ariely & Loewenstein, 2000). Journal of Experimental Psychology: General. 2000; 129:524–529. PubMed. [PubMed: 11142866]

Ariely D, Loewenstein G. When does duration matter in judgment and decision making? Journal of Experimental Psychology: General. 2000; 129:508–523. PubMed. [PubMed: 11142865]

Bar M, Neta M. Humans prefer curved visual objects. Psychological Science. 2006; 17:645–648. PubMed. [PubMed: 16913943]

Berlyne DE. The influence of complexity and novelty in visual figures on orienting responses. Journal of Experimental Psychology. 1958; 55:289–296. PubMed. [PubMed: 13513951]

Berlyne DE. Novelty, complexity, and hedonic value. Perception & Psychophysics. 1970; 8:279–286.

Berlyne, DE. Aesthetics and psychobiology. Meredith Corporation; New York: 1971.

Berlyne, DE.; Ogilvie, JC. Dimensions of perception of paintings. In: Berlyne, DE., editor. Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation. Hemisphere Publishing Corporation; Washington, D.C.: 1974. p. 181-226.

Biederman I, Vessel EA. Perceptual pleasure and the brain. American Scientist. 2006; 94:247–253.

Bornstein RF. Exposure and affect—Overview and meta-analysis of research, 1968-1987. Psychological Bulletin. 1989; 106:265–289.

Bradley RA, Terry ME. Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika. 1952; 39:324–345.

Brainard DH. The Psychophysics Toolbox. Spatial Vision. 1997; 10:433–436. PubMed. [PubMed: 9176952]

Bronstad PM, Russell R. Beauty is in the ‘we’ of the beholder: Greater agreement on facial attractiveness among close relations. Perception. 2007; 36:1674–1681. PubMed. [PubMed: 18265847]

Cupchik, GC. An experimental investigation of perceptual and stylistic dimensions of paintings suggested by art history. In: Berlyne, DE., editor. Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation. Hemisphere Publishing Corporation; Washington, D.C.: 1974. p. 235-257.

Cupchik GC, Gebotys RJ. Interest and pleasure as dimensions of aesthetic response. Empirical Studies of the Arts. 1990; 8:1–14.

Ellsworth, PC.; Scherer, KR. Appraisal processes in emotion. In: Davidson, RJ.; Scherer, KR.; Goldsmith, H., editors. Handbook of affective sciences. Oxford University Press; New York: 2003. p. 572-595.

Eysenck HJ. The general factor in aesthetic judgments. British Journal of Psychology. 1940; 31:94–102.

Eysenck, HJ. Genius: The natural history of creativity. Cambridge University Press; Cambridge, UK: 1995.

Fechner, GT. Elemente der Psychophysik (Trans. H.S. Langfield). In: Rand, B., editor. The Classical Psychologists. Houghton Mifflin; Boston: 1912. p. 562-572.Originally published 1860

Fredrickson BL, Kahneman D. Duration neglect in retrospective evaluations of affective episodes. Journal of Personality and Social Psychology. 1993; 65:45–55. PubMed. [PubMed: 8355141]

Gangestad SW, Thornhill R, Yeo RA. Facial attractiveness, developmental stability, and fluctuating asymmetry. Ethology and Sociobiology. 1994; 15:73–85.

Graham DJ, Field DJ. Statistical regularities of art images and natural scenes: Spectra, sparseness and nonlinearities. Spatial Vision. 2007; 21:149–164. PubMed. [PubMed: 18073056]

Hönekopp J. Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. Journal of Experimental Psychology: Human Perception and Performance. 2006; 32:199–209. PubMed. [PubMed: 16634665]

Kaplan, R.; Kaplan, S. The experience of nature: A psychological perspective. Ulrich's Bookstore; Ann Arbor, Michigan: 1995.

Kaplan, S. Environmental preference in a knowledge-seeking, knowledge-using organism. In: Barkow, JH.; Cosmides, L.; Tooby, J., editors. The adapted mind: Evolutionary psychology and the generation of culture. Oxford; New York: 1992. p. 581-598.

Leder H, Belke B, Oeberst A, Augustin D. A model of aesthetic appreciation and aesthetic judgments. British Journal of Psychology. 2004; 95:489–508. PubMed. [PubMed: 15527534]

Luce, RD. Individual choice behavior. Wiley; New York: 1959.

Martindale C. The pleasures of thought: A theory of cognitive hedonics. The Journal of Mind and Behavior. 1984; 5:49–80.

Martindale C, Moore K, Borkum J. Aesthetic preference—Anomalous findings for Berlyne's psychobiological theory. American Journal of Psychology. 1990; 103:53–80.

McGuire DP, Davison ML. Testing group-differences in paired comparisons data. Psychological Bulletin. 1991; 110:171–182.

McManus IC. The aesthetics of simple figures. British Journal of Psychology. 1980; 71:505–524. PubMed. [PubMed: 7437674]

McManus IC, Jones AL, Cottrell J. The aesthetics of colour. Perception. 1981; 10:651–666. PubMed. [PubMed: 7110879]

Millis K. Making meaning brings pleasure: The influence of titles on aesthetic experiences. Emotion. 2001; 1:320–329. PubMed. [PubMed: 12934689]

Nijenhuis, A.; Wilf, HS. Combinatorial algorithms for computers and calculators. 2 ed.. Academic Press; New York: 1978.

Orians, GH.; Heerwagen, JH. Evolved responses to landscapes. In: Barkow, JH.; Cosmides, L.; Tooby, J., editors. The adapted mind: Evolutionary psychology and the generation of culture. Oxford University Press; New York: 1992. p. 555-579.

Parducci A, Knobel S, Thomas C. Independent contexts for category ratings: A range-frequency analysis. Perception & Psychophysics. 1976; 20:360–366.

Parducci A, Wedell DH. The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. Journal of Experimental Psychology: Human Perception and Performance. 1986; 12:496–516. PubMed. [PubMed: 2946806]

Pelli DG. The videotoolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision. 1997; 10:437–442. PubMed. [PubMed: 9176953]

Rawlings D. The interaction of openness to experience and schizotypy in predicting preference for abstract and violent paintings. Empirical Studies of the Arts. 2000; 18:69–91.

Reber R, Schwarz N, Winkielman P. Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? Personality and Social Psychology Review. 2004; 8:364–382. PubMed. [PubMed: 15582859]

Reber R, Winkielman P, Schwarz N. Effects of perceptual fluency on affective judgments. Psychological Science. 1998; 9:45–48.

Rentschler I, Jüttner M, Unzicker A, Landis T. Innate and learned components of human visual preference. Current Biology. 1999; 9:665–671. PubMed. [PubMed: 10395537]

Rhodes G, Sumich A, Byatt G. Are average facial configurations attractive only because of their symmetry? Psychological Science. 1999; 10:52–58.

Russell PA. Effort after meaning and the hedonic value of paintings. British Journal of Psychology. 2003; 94:99–110. PubMed. [PubMed: 12648392]

Shepard RN. Recognition memory for words, sentences, and pictures. Journal of Verbal Learning and Verbal Behavior. 1967; 6:156–163.

Shimojo S, Simion C, Shimojo E, Scheier C. Gaze bias both reflects and influences preference. Nature Neuroscience. 2003; 6:1317–1322. PubMed.

Shortess GK, Clarke CJ, Richter ML, Seay M. Abstract or realistic? Prototypicality of paintings. Visual Arts Research. 2000; 26:70–79.

Silvia PJ. Cognitive appraisals and interest in visual art: Exploring an appraisal theory of aesthetic emotions. Empirical Studies of the Arts. 2005a; 23:119–133.

Silvia PJ. Emotional responses to art: From collation and arousal to cognition and emotion. Review of General Psychology. 2005b; 9:342–357.

Standing L. Learning 10,000 pictures. Quarterly Journal of Experimental Psychology. 1973; 25:207–222. PubMed. [PubMed: 4515818]

Thornhill R, Gangestad SW. Facial attractiveness. Trends in Cognitive Sciences. 1999; 3:452–460. PubMed. [PubMed: 10562724]

Tinio PPL, Leder H. Natural scenes are indeed preferred, but image quality might have the last word. Psychology of Aesthetics, Creativity, and the Arts. 2009; 3:52–56.

Van Tonder GJ, Lyons MJ, Ejima Y. Visual structure of a Japanese Zen garden. Nature. 2002; 419:359–360. PubMed. [PubMed: 12353024]

Vessel EA, Biederman I. An fMRI investigation of visual preference habituation [Abstract]. Journal of Vision. 2002; 2(7):492. 492a, http://journalofvision.org/2/7/492/, doi:10.1167/2.7.492.

Weber, EH. Der Tastsinn and das Gemeingefuhl. In: Wagner, R., editor. HandwOrterbuch der Physiologie. Vol. iii. 1846.

Williams JWJ. Algorithm-232—Heapsort. Communications of the ACM. 1964; 7:347–348.

Wilson, EO. Biophilia and the conservation ethic. In: Kellert, SL.; Wilson, EO., editors. The Biophilia hypothesis. Island Press; Covelo, California: 1993. p. 31-41.

Winkielman P, Cacioppo JT. Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation elicits positive affect. Journal of Personality and Social Psychology. 2001; 81:989–1000. PubMed. [PubMed: 11761320]

Yu DW, Shepard GH. Is beauty in the eye of the beholder? Nature. 1998; 396:321–322. PubMed. [PubMed: 9845067]

Yue X, Vessel EA, Biederman I. The neural basis of scene preferences. NeuroReport. 2007; 18:525–529. PubMed. [PubMed: 17413651]

Zajonc RB. Attitudinal effects of mere exposure. Journal of Personality and Social Psychology. 1968; 9:1–27. Monograph supplement No. 2, Part 2. [PubMed: 5667435]

Zuckerman M, Ulrich RS, McLaughlin J. Sensation seeking and reactions to nature paintings. Personality and Individual Differences. 1993; 15:563–576.

Figure 1.
Examples of the real-world (top) and abstract (bottom) images used. (a) Previous work has found high agreement across observers in visual preferences for real-world scenes such as these. From left to right, images of high, medium, and low preference. (b) The abstract images created for this experiment. See the Supplementary Material for the entire stimulus set.

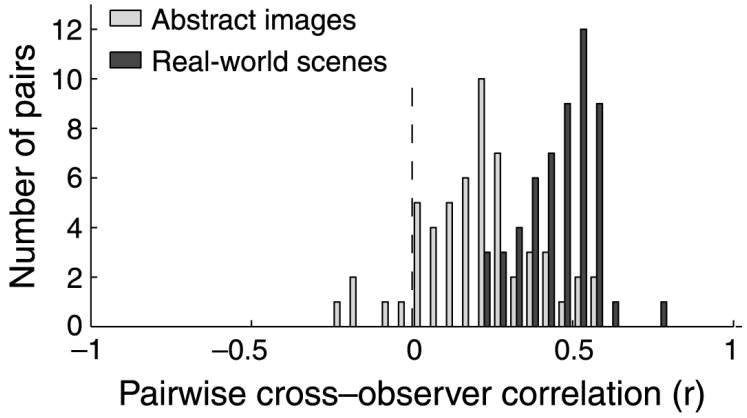

Figure 2.
Pairwise correlations between observers of their visual preferences when the two image sets were judged in separate sessions (Experiment 1). Across-observer correlations were significantly higher for real-world scenes than for abstract images.

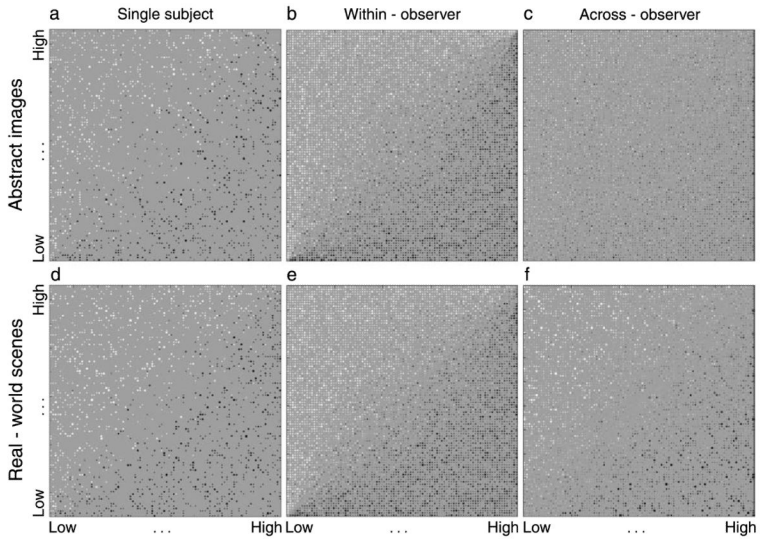

Figure 3.

Choice matrices both for a single subject and collapsed across observers. The grayscale value of each tile indicates the proportion of trials at which the stimulus on the vertical axis was preferred over the stimulus on the horizontal axis (dark to bright indicating 0 to 1). The area of each tile indicates the number of comparisons made between the two stimuli, and is therefore an indicator of confidence. Panels (a) and (d) show choice matrices for a single observer for abstract images and real-world scenes, respectively. They were sorted based on the preferences for this observer, whom had the median difference between reliability scores for the two image types. Panels (b) and (e) show the within-observer reliability matrices for abstract images and real-world scenes, respectively. They were created by sorting each observer's choice matrix by his/her *own* preferences before collapsing across observers. Panels (c) and (f) show the across-observer agreement matrices for abstract images and real-world scenes, respectively. They were created by sorting each observer's choice matrix by the *group* average preferences before collapsing across observers. The coherence of the within-observer reliability matrices is high for both types of stimuli (b and e). In contrast, the coherence for the across-observer agreement matrices is strikingly lower for abstract images (c) than for real-world scenes (f).

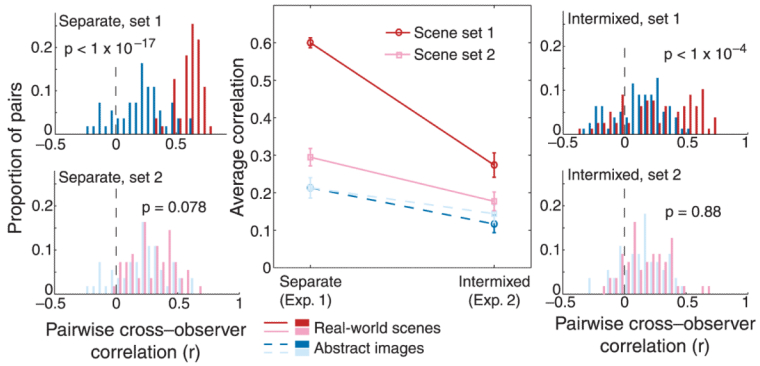

Figure 4.

Effect of intermixing abstract images and real-world scenes. The central panel shows the average across-observer correlation for real-world scenes (red) and abstract images (blue) when judged in separate sessions (Experiment 1) versus when these stimuli were intermixed in the same session (Experiment 2). Experiment 2 was run with two different sets of real-world scenes, consisting of subsets of the stimuli from Experiment 1 that yielded higher versus lower than average agreements (Sets 1 and 2, respectively). Both sets showed significant drops in across-observer agreement when intermixed with abstract stimuli, with Set 1 showing the more dramatic drop. The panels on the left and right show the full distributions of across-observer pairwise correlations in the separate (Experiment 1) and intermixed (Experiment 2) conditions, respectively.