Finding the "odd one out": Memory color effects and the logic of appearance

J.J. Valenti and Chaz Firestone

Johns Hopkins University

Header: Memory Color And Perceptual Logic
Version: 4/3/19 — in press, *Cognition*
Word Count: 10920

Address for correspondence:
Chaz Firestone (*chaz@jhu.edu*)
Department of Psychological & Brain Sciences
Johns Hopkins University
3400 N Charles St
Baltimore, Maryland 21218

Abstract

Can what we know change what we see? A line of research stretching back nearly a century suggests that knowing an object's canonical color can alter its visual appearance, such that objectively gray bananas appear to be tinged with yellow, and objectively orange hearts appear redder than they really are. Such "memory color" effects have constituted the strongest and most complete evidence that basic sensory processing can be penetrated by higher-level knowledge, and have contributed to theories of object perception in psychology, neuroscience, and philosophy. Are such phenomena truly perceptual? Or could they instead reflect shifts in judgments and responses without altering online color perception? Here, we take a novel approach to this question by exploiting a "logic" that is inherent in visual processing but that higher-level cognition often cannot follow. In four experiments spanning both classical and contemporary work, we exhaust the predictions of memory color theories, by exploring scenarios where memory color accounts make tortuous and difficult-to-grasp hypotheses that should nevertheless be easily accommodated by visual processing. We show that such conditions eliminate or even reverse memory color effects in ways unaccounted-for by their underlying theories — especially in a novel "odd one out" paradigm that may help distinguish visual appearance from higher-level judgment in a powerful and general way. We suggest that prior knowledge can influence color judgments in real and robust ways, but that such influences may not truly reflect changes in visual appearance per se. We further discuss the general utility of this approach for isolating perception from judgment, both for memory color effects and beyond.

Keywords: memory color, color perception, top-down effects, cognitive penetrability, modularity

What we see can change what we believe — but can what we believe change what we see? In contrast to a traditional "modular" view of perception, recent work has suggested that higher-level cognitive states can reach down into visual processing and change how the world appears to us. For example, it has been reported that recalling unethical behavior can change the perceived brightness of a room (Banerjee, Chatterjee, & Sinha, 2012), that impressive or powerful people appear larger (Duguid & Goncalo, 2012; Masters et al., 2010), and that frightening objects appear closer (Harber et al., 2011). Such work goes beyond the more widely accepted notion that higher-level expectations can modulate attention, eye movements, or object categorization (Bar, 2004; Malcolm et al., 2014); instead, the force of these results is their claim to directly alter the way a given object *looks* to us, at the level of basic visual properties such as color, size, or distance.

The ground-shaking consequences implied by these claims have made them extremely influential, but they have also attracted skepticism for the same reasons, on at least three general fronts. First, the studies used to motivate these claims are often "one-off" results rather than integrated parts of a broader literature: With rare exceptions (e.g., Dunning & Balcetis, 2013; Proffitt, 2006; Witt, 2011; but see also Durgin et al., 2011; Firestone 2013a), many of these claims draw from only one source of evidence, use only one sort of task, and have rarely been replicated or extended by other researchers and laboratories. Second, these studies frequently arise from outside the field of vision science, and as a result may overlook certain controls expected of perception studies, such as matching stimuli on important low-level properties (e.g., Harber et al., 2011; van Ulzen et al., 2008). Third, many of these claims are theoretically puzzling and even implausibly maladaptive. For example, given the degree to which visual processing relies on representing a scene's lighting conditions, it would seem odd at best (and actively unhelpful at worst) for the mind to adjust a room's perceived brightness when the perceiver has recently recalled an unethical vs. ethical action (Banerjee et al., 2012) — an effect that has no clear adaptive function and could even mislead perceivers about their visual environment. These and other concerns have motivated reviewers of this literature to ask whether there is truly any evidence that cognition can penetrate visual perception (Firestone & Scholl, 2016; Lammers, de Haan, & Pinto, 2017; Machery, 2015; for more classical discussions of modularity and cognitive impenetrability; see Fodor, 1983; Pylyshyn, 1999).

The highest-hanging fruit

In the entire literature on top-down effects of cognition on perception, one class of findings stands apart in straightforwardly overcoming many of the above weaknesses: a collection of results known as "memory color" effects. Memory color effects are said to occur when one's prior belief about the color of an object changes the color one actually experiences that object to be. For example, an orange-red heart may appear redder than it really is (Delk & Fillenbaum, 1965), or a gray banana may appear tinged with yellow (Hansen et al., 2006; Olkkonen et al., 2008), because the perceiver knows the objects' canonical colors (Figure 1).

Memory color effects withstand many of the criticisms raised against other top-down effects of cognition on perception. First, they have been observed for nearly a century and have been replicated in various ways since their initial discovery (e.g., Adams, 1923; Bannert & Bartels, 2013; Bruner, Postman, & Rodrigues, 1951; Delk & Fillenbaum, 1965; Duncker, 1939; Hansen et al., 2006; Lupyan, 2015b; White & Montgomery, 1976; Witzel, 2016; for a review, see Adeyefa-Olasupo & Flombaum, 2018). Second, though some classical memory color effects may have had methodological shortcomings by today's standards, the research program has been revived by modern vision science laboratories that pay careful attention to many important methodological details, including various controls for the stimuli used in the experiments, the manner in which the measurements are taken, and experimenter bias (e.g., Hansen et al., 2006; Olkkonen et al., 2008; Witzel et al., 2011; Witzel, 2016). Third, the findings *make sense* in a way that so many other alleged top-down effects on perception do not: If the perceiver knows something about the typical color of an object, then incorporating that prior information into the visual system's computation of that object's color doesn't seem so unreasonable (and may perhaps be consistent with 'Bayesian' norms of inference; Witzel, Olkkonen, & Gegenfurtner, 2018).

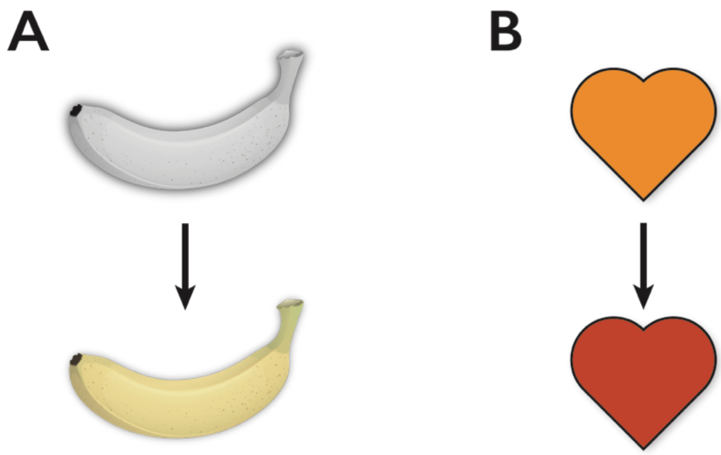

Figure 1. Schematic illustration (and exaggeration) of memory color effects. (A) Objectively gray bananas may appear tinged with yellow. (B) Hearts may appear redder than they really are.

For these reasons and others, the memory color effect is considered by many researchers to be among the most promising candidates for a genuine top-down effect of cognition on perception. Indeed, effects of knowledge on color appearance have played a central role in recent arguments for the cognitive penetrability of perception (e.g., Macpherson, 2012; Newen & Vetter, 2017; Vetter & Newen, 2014), have helped to motivate new perspectives on cognitive architecture more generally (e.g., Barsalou, 2008; Lupyan, 2015a), and have even appeared in popular perception textbooks (e.g., Goldstein & Brockmole, 2016; Schwartz & Krantz, 2017).

A seed of doubt?

At the same time that memory color effects have been so robust and influential, there is also reason to wonder whether they truly reflect changes in *perception* — how an object *looks*, per se — vs. changes in subjects' behavioral responses, which may occur without altering what is actually seen. In other words, might it possible that subjects give *responses* that are in line with memory color theories, even if they don't literally *see* objects as having those colors? In particular, one might doubt a perceptual interpretation of these effects for at least three reasons:

Effect sizes. First, memory color effects are surprisingly large — so large that they should be subjectively apparent even when casually viewing the experimental stimuli (as in Figure 1a, where memory color theory predicts that the gray banana shown there should look yellow to *you*, the reader, right now). For example, many modern studies of memory color effects measure their presence using the method of "achromatic adjustment", whereby subjects adjust a stimulus's color until it appears to be a neutral gray (e.g., Hansen et al., 2006; Lupyan, 2015b; Olkkonen et al., 2008; Witzel et al., 2011). In the case of a yellow banana, for example, subjects in fact make the banana a bit *blue* — with the idea being that the visual system is independently adding some extra yellow to the image's perceived color, such that the banana image must be objectively blue (yellow's opponent color) in order to cancel out the extra yellow and thereby appear gray to the subject. What is striking about such effects is just *how much* color subjects add to the images: In the original studies, the effects were as high as 13% relative to the images' typical color settings (with a mean of 8%; Hansen et al., 2006), and follow-up work has adjusted this value even higher, estimating the memory color effect for grayscale photographs of bananas to be 22% (Olkkonen et al., 2008). This essentially implies that an objectively gray banana should appear to be 22% as colorful as a real-life, naturally colored yellow banana. Hansen et al. note that a finding of this magnitude "amounts to an effect that is approximately three to five times above the threshold of discrimination" (p.1368); but on reflection this seems like a *problem*, because an effect of this size just does not comport with the subjective experience of seeing a gray banana — which, fairly plainly, looks gray (as in Figure 1a). In other words, speaking purely subjectively, gray bananas just don't *look* yellow to the degree suggested by memory color studies. (For a more sustained treatment of the role of phenomenology in evaluating top-down effects, see Firestone & Scholl, 2015a.)

Conflicting mechanisms. Second, there may be inconsistencies across memory color studies as to their underlying mechanisms — in particular, as to whether memory color effects are driven by explicit higher-level beliefs or instead by low-level statistical associations between shapes and colors. For example, it has been reported that blurring a fruit image reduces its memory color effect even when subjects still know the fruit's identity (Olkkonen et al., 2008). This suggests a fairly low-level learning mechanism for memory color effects, since even when the subject's high-level knowledge is preserved (i.e., even when the subject still knows they are viewing a banana), the blurring procedure reduces the memory color effect (see Olkkonen et al., 2008, for a more detailed interpretation of this finding along these lines). At the same time, however, the largest memory color effects of all are typically observed for branded objects with very generic shapes, where the effect seems to be driven by exactly the sort of higher-level knowledge that earlier studies seemed to have ruled out. For example, a simple disk does not produce a memory color effect on its own; but when the disk is branded with the logo for Nivea (a popular brand of skin lotion with dark blue containers), a large memory color effect occurs, such that the additional writing on the surface of the tin reportedly causes the whole tin to appear blue (Witzel et al., 2011). This result suggests a higher-level source of memory color effects (what those researchers call "object knowledge"), since a rounded tin is a very generic shape, and it's only the addition of meaningful writing that now produces a large memory color effect. Though these somewhat conflicting patterns do not by themselves undermine the relevant studies, they may well be a reason to seek alternative explanations for many of these effects.

Unconstrained tasks. Third, the vast majority of contemporary memory color studies use tasks that rely on the subjects' pre-existing notions of where "gray" and other colors are located in color space, in a potentially problematic way. For example, rather than show subjects a gray object and ask them whether a gray banana *looks like that*, most memory color studies show subjects a colored banana and ask subjects to *make it gray*, leaving it to the subjects themselves to determine the gray standard on their own. This open-ended design may be especially susceptible to biases in judgments and responses, especially since there are usually no countermeasures in place to prevent subjects from inferring the purpose of these studies (which, given their designs, plainly concern the connection between familiar objects and their typical colors). Just like any color term, "gray" does not correspond to a single point in color space, but instead to a continuous region within that space: As one discovers in a paint store, for example, many different colors answer to "gray", including not only dark grays and light grays but also warm grays and cool grays, which have other hues mixed in but are still accepted as instances of "gray". Just as asking a subject to adjust something to be "red" leaves open the precise shade of red they should choose — and even gives the subject leeway to choose different reds for different objects depending on each object's canonical red — asking a subject to make something "gray" permits the subject to choose a different gray for different objects, or to use different strategies for different objects, even if all such objects are *perceived* identically (and if that subject's own subjective gray standard is measured in advance). Indeed, one distinct possibility is that subjects who adjust bananas to be a bit blue do so because their strategy is to stay a safe distance away from the object's canonical color — as if thinking, "*Let me make certain there's not even a speck of yellow in this banana*" — and so tend to overshoot towards that color's opponent (here, blue). In other words, if subjects wish to make a banana look gray, and in so doing make a special effort to avoid yellowish grays, then a banana might elicit a bluish estimate for strategic reasons ("*let me reduce yellow a bit more, just to be safe*") rather than because of how the banana visually appears.[1] (For discussion of a similar worry, see Zeimbekis, 2013.)

[1] Indeed, if subjects take this approach to the achromatic adjustment task, then this might even explain why memory color effects revealed by that task have typically been stronger for colors on the daylight axis (e.g., blues and yellows) than for other colors, especially reds (e.g., a strawberry image in Olkkonen et al., 2008; though see Delk & Fillenbaum, 1965, who do find memory color effects for red objects). Since color discrimination is poorer along the blue-yellow

Of course, these issues do not by themselves show that memory color effects are not perceptual, and they certainly do not undermine the reality or robustness of such effects. But they do raise the question: Might such effects not reflect how a given object visually *appears* in the moment, but instead how the subject *responds* in certain experimental conditions?

Separating seeing from thinking through the "logic of appearance"

How could we distinguish biases in *perception* from biases in post-perceptual *judgment*? Our approach here is to take advantage of a kind of "logic" that is automatically employed by visual processing (Rock, 1983) but that operates more slowly and less reliably (if at all) in higher-level cognition. In particular, truly perceptual phenomena typically arise not only in reports about the stimuli undergoing those phenomena, but also in our experience of how those stimuli *relate* to other objects and events in the environment.

For example, ask yourself: "If a teal object had its color move roughly half the distance towards red in color space, and another teal object had its color move roughly half the distance towards green, and they both appeared beside a light purple object, which of the three objects would look the most different from the other two?" While you find yourself working through this problem in your head, look at Figure 2 to *see* the answer for yourself. Figure 2a shows two objectively identical teal ovals (center) whose appearance has been altered by the Munker illusion: The oval on the left looks purplish, and the oval on the right looks greenish — and if subjects were asked to explicitly report the colors of these ovals, they would surely say so. Additionally, however, the same sort of effect can be revealed by *comparison*: When the two ovals appear next to an objectively purple oval (Figure 2b), the greenish-looking oval now looks *different* from the other two. If you were asked to identify the "odd color out" from the display, you could easily point to the rightmost, greenish-looking oval.

---

daylight axis (Pearce et al., 2014), it may be easier to discriminate purely achromatic gray from a slightly greenish gray than it is to discriminate purely achromatic gray from a slightly bluish gray. In that case, subjects who are asked to make a banana "gray" are able to stray relatively far into blue territory while still judging the blue-gray sample to be acceptably gray; by contrast, subjects who are asked to make a strawberry "gray" will tolerate only just a bit of green in the sample, because anything more would be easily noticed. The relative weakness or absence of memory color effects for red objects has sometimes been taken as evidence for their perceptual nature (e.g., Block, 2016); but in fact, even this fairly specific pattern of results can still be explained by biased responding, once other baseline aspects of perceptual discrimination are taken into account (in particular, baseline differences in the discriminability of different colors).

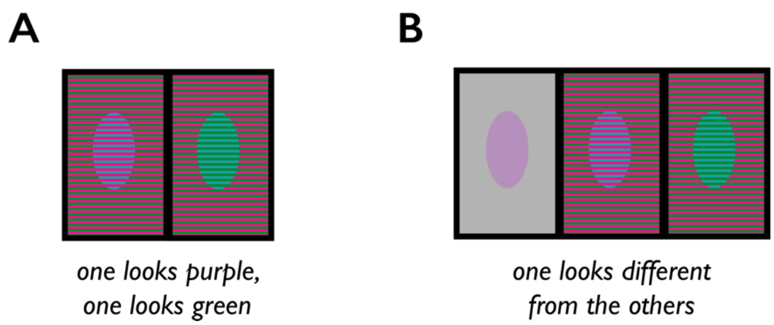

Figure 2. An example of "logic" in perceptual comparison. (A) Objectively, the two ovals are identical shades of teal, but the Munker illusion causes the left one to appear purplish and the right one to appear greenish. (B) Next to an objectively purple oval, the green-looking oval not only looks green but also appears to the "odd one out" in the triplet in terms of its hue. Note that this is true even without a *perfect* hue match; even if the two purplish-looking ovals in panel B don't look identical, they still may look similar enough to seem different than the greenish-looking one.

Notice that this straightforward experience of one object standing out from the rest reflects a kind of "logical" relationship among otherwise-basic perceptual processes: First, multiple stimuli undergo the Munker illusion, in which a stimulus is assimilated to the color of its foreground; then, the stimuli that now carry these illusory colors are each compared to one another and to a third stimulus, which looks as it does for independent reasons; finally, one of these pairwise comparisons comes out "same" while two of them come out "different", resulting in one object standing out from the rest. When reading the riddle-like question that began the previous paragraph, one finds oneself laboring through this logic, as if slowly and tentatively solving a puzzle; but that same logic is implemented flawlessly and near-instantaneously in visual processing, as you can see in Figures 2b and 2c.

Crucially for our purposes here, note that the "odd one out" judgments you can make in Figures 2b and 2c never require you to actually *name* the color that the objects are. Indeed, a distinct advantage of this approach is that one can obtain evidence that the illusion is working without ever asking which color the objects appear to be; given the logic of these comparisons, all that is required is a judgment about which object looks most different from the others.

The present studies: The logic of appearance in memory color effects

Here, we exploit this perceptual logic for the case of memory color effects. Rather than ask subjects to report the colors they see, and rather than make subjects apply a pre-conceived color standard, we show subjects *sets* of objects that have or don't have canonical colors, and we ask subjects to identify similarities and differences between them. Experiments 1 and 2 apply this multi-step logic of comparison to contemporary memory color effects discovered in recent years, while Experiments 3 and 4 explore classical memory color effects from the middle of the last century. In all cases, we ask whether such effects persist when their underlying theories make predictions that would be difficult for subjects to follow or provide strategic answers for.

In other words, we ask: Do memory color effects obey the logic of appearance?

Experiment 1: The "odd-one-out" task

If memory color effects are genuinely perceptual, then they should reveal themselves not only through direct reports of an object's color, but also through "odd-one-out" comparisons of the sort in Figure 2. (For other perception studies employing analogous tasks in other domains, see Adams, Kerrigan, & Graf, 2016; Robilotto & Zaidi, 2004.) Experiment 1 ran such a test for the now-classic case of yellow-looking bananas.

Recent work has established that the memory color effect for such stimuli can be measured not only through achromatic adjustment (as in Hansen et al., 2006), but also through a much simpler forced-choice task in which subjects are shown an objectively gray banana next to an objectively bluish banana whose color matches the particular shade of blue that independent subjects selected when making a banana gray. When asked which of the two objects is truly "gray", subjects tend to select the objectively bluish banana more often than the objectively gray banana — however, subjects do not do this for bluish and gray *disks* (Witzel, 2016). The interpretation of this pattern was that, due to memory color effects, blue bananas look gray and gray bananas look yellow, such that subjects select the bluish banana as "gray"; but since disks don't have canonical colors, there is no similar effect in those cases.

This method, helpfully named the "Easy Way to Show Memory Color Effects" (by Witzel, 2016), does indeed have several advantages over earlier methods. Notably, its design and key comparisons mean that the experiment does not depend on an elaborate setup with sensitive equipment, precise calibration, and a lengthy experimental session (as earlier investigations of memory color have required); indeed, this new forced-choice method was run over the Internet, using whatever display the subjects happened to have at home. This advance allows for a straightforward and highly standardized design that can be easily replicated across laboratories, and also permits much larger samples than have been used in previous studies, given the ease of online data collection. Partly for these reasons, this new approach involving direct comparison between objects has motivated especially strong claims about separating perception from judgment, as it has been explicitly argued that such experiments "measure a bias in perception rather than memory or judgment biases" (Witzel & Gegenfurtner, 2018, p.480).

Here, we make a simple change to the forced-choice task used in such experiments: Rather than select the "gray" object among two choices, we ask subjects to select the "odd color out" among *three* choices. For example, we ask subjects to select which object among a bluish disk, a bluish banana, and a gray disk appears to be a different color than the other two (Figure 3). If bluish bananas of this particular blue tend to look gray (as predicted by memory color theories and reported in previous work), then the bluish banana and the gray disk should look similar, leaving the bluish disk appearing to be the odd color out among the three. But if subjects simply see the colors roughly as they are (as predicted by modular theories), then they should select the gray disk as the odd color out.

Method

Subjects

220 subjects were recruited online from Amazon Mechanical-Turk and were monetarily reimbursed. (For discussion of this subject pool's reliability, see Crump et al., 2013; for other investigations of color perception run online, see Haberman, Brady, & Alvarez, 2015; Lafer-Sousa, Hermann, & Conway, 2015; Witzel, 2016; Witzel et al., 2017). This sample size followed Witzel (2016, Study 2), which ran 210 subjects; we slightly increased this number (both in this study and in Experiment 2) because of an "attention check" we added that we expected would lead to slightly more exclusions than in previous work.

Stimuli

The same four images as used in Witzel (2016) served as stimuli in the present experiment: a gray disk, a gray banana, a bluish disk, and a bluish banana, presented on a neutral gray background.[2] The particular blue of the bluish objects was used in previous work because it approximates the

[2] We thank Christoph Witzel for generously providing these stimuli.

color that subjects adjust a banana to be when asked to make it "gray"; importantly, this previous work showed that subjects consider the bluish banana to be a better example of "gray" than even the objectively gray banana.

From these four images, we created four trial types corresponding to the four unique triplets that can be made from them. Each triplet comprised three side-by-side images, displayed in a newly randomized left-to-right order each time they were presented (for a "screenshot" of the task, see Figure 3). Due to the nature of online data collection, we cannot be sure of the exact color, size, or brightness (etc.) of the images as they appeared to subjects during the experiment; however, any distortions or miscalibrations caused by a given subject's monitor would have been held constant across conditions (just as in earlier online studies using these stimuli).

In addition to these experimental trials, the final trial of the experiment was a "catch trial" consisting of a triplet that included a green square, a blue square, and a blue circle whose colors were much more saturated than those of the experimental stimuli; the purpose of this trial was to ensure that the subjects were paying attention (see below).

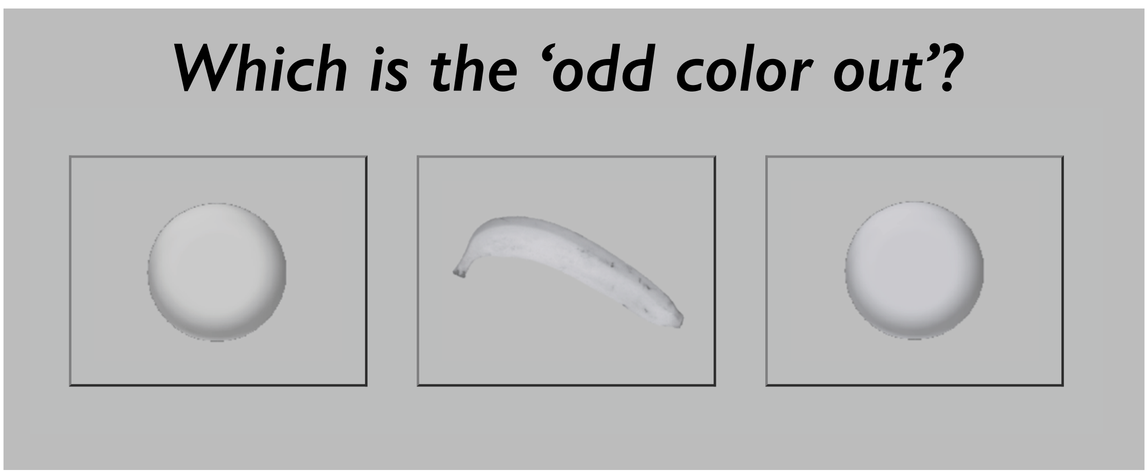

Figure 3. An example triplet from Experiment 1. On each trial, subjects saw three images in a random order, and selected whichever image appeared to have a different color than the other two. This particular trial shows a gray disk (left), blue banana (middle), and blue disk (right), and so the objectively "correct" answer would be to select the left-most image. By contrast, memory color theories predict that the blue disk (right) should appear to be the odd color out (or, at least, the most different in hue), since the blue banana has exactly the right amount of blue to appear gray (as measured and validated in previous studies).

Procedure

Subjects were first shown an instruction page which informed them that they would see three objects on each of many subsequent pages, and that they should "judge which one looks to be a slightly different color from the other two, by clicking on that object." The instructions page also showed subjects an "easy" practice triplet that included a (much more saturated) blue circle, yellow square, and yellow circle; subjects were told, for this example, that they "should select the *blue circle*, since it is a different color from the other two objects", but also that "the real experiment will be much harder than this, though! The colors will be very hard to tell apart, so look closely."

In the experiment itself, subjects completed the odd-color-out task 10 times for each of the four trial types, for a total of 40 experimental trials; the trials appeared in blocks of four (one for each of the four trial types), with the trial order randomized within each block. The relative position of each individual image within the triplet (i.e. left, middle, or right) was randomly chosen for each trial. Subjects could only advance to the next trial after clicking on one of the three images displayed. There was no time limit on responses. After making a selection, all of the images disappeared from the screen, followed by a 2000ms interval before the next trial's images appeared.

After all 40 experimental trials, the "catch" trial appeared; this was later used as an exclusion criterion to ensure that subjects understood the task.

Readers can experience the task for themselves at http://perceptionresearch.org/bananas.

Results and Discussion

21 subjects were excluded either for failing to provide a complete dataset (4/220) or for failing to correctly answer the "catch" question (17/220), leaving 199 subjects with usable data. However, none of the results reported here depended on these exclusions (i.e. all of the effects below remain statistically significant, in the same direction, even without excluding these subjects).

For each triplet type, we can consider the prediction made by the memory color view (according to which blue bananas appear gray, and gray bananas appear yellow) and the prediction made by the "modular" view (according to which color knowledge does not affect color appearance), and compare the data to those predictions.[3] To foreshadow the general pattern, every triplet yielded the result predicted by the modular view, and none of them yielded the result predicted by the memory color view (Figure 4).

[3] Given that there were 10 repetitions of each triplet, it was possible for subjects to respond inconsistently across repetitions, which may have been a marker of low engagement on the part of the subject; to ensure that random or unthoughtful responses did not contaminate the results, we considered responses only from those subjects who consistently selected the same triplet-member a majority of the time across repetitions (i.e. >5 times out of 10 opportunities). This left 148 subjects (from the original 199). Once again, however, none of the results depended on such exclusions: All of the contrasts and inferential statistics reported here remain statistically significant even when these subjects are not excluded.

For the triplet consisting of {*gray disk, bluish banana, bluish disk*}, memory color theory predicts that the blue banana should appear gray, and thus that subjects should pick the bluish disk as the odd color out; by contrast, the modular view predicts that subjects should pick the gray disk as the odd color out, since the bluish banana should look similar to the equally blue disk (Figure 4a). In fact, 68.9% of subjects selected the gray disk as the odd color out (the choice consistent with the modular view), and only 1.4% of subjects selected the blue disk as the odd color out (the choice consistent with memory color), $\chi^2(2, N=148)=102.32$, $p<.001$; the remaining 29.7% of subjects selected the bluish banana as the odd color out, which is predicted by neither view. (Given the subtlety of the differences in color for these images, we suspect that subjects who couldn't detect a meaningful difference in color between the three images simply defaulted to picking the odd *shape* out, which in this triplet was the bluish banana. Another possibility is that, being unable to detect any difference in hue, subjects considered the difference in shading of this third object to be a relevant difference in "color", and so chose it for that reason. As is clear below, other conditions reveal a similar odd-shape-out pattern.)

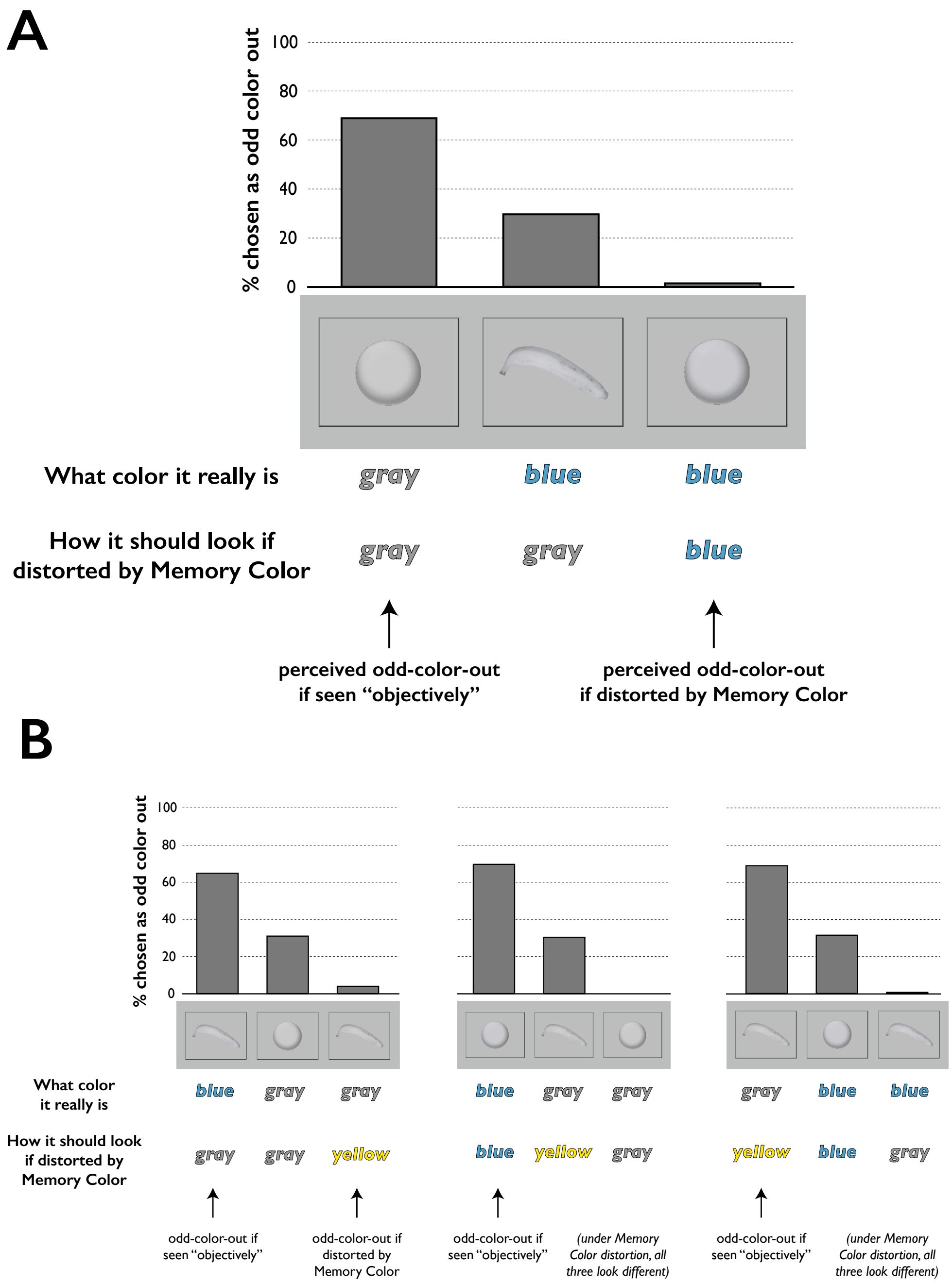

Figure 4. Results from Experiment 1. (A) Subjects accurately identified a gray disk as the "odd color out", even though memory color theories predict that the blue disk should be perceived as the odd color out (because the blue banana should appear gray). (B) Across all trial types, the most popular selection was always the perceived odd color out predicted by a modular view, and never the perceived odd color out predicted by memory color theory.

For the triplet consisting of {*bluish banana, gray disk, gray banana*}, memory color theory predicts that the blue banana should appear gray (like the gray disk), and the gray banana should appear yellow, and thus that subjects should pick the yellow-looking gray banana as the odd color out; by contrast, the modular view predicts that subjects should pick the bluish banana as the odd color out, since the gray banana should look similar to the equally gray disk, and the bluish banana should look blue. In fact, 64.8% of subjects selected the bluish banana as the odd color out (the choice consistent with the modular view), and only 4.1% of subjects selected the gray banana as the odd color out (the choice consistent with memory color); $\chi^2(2, N=148)=82.52$, $p<.001$. The remaining 31.1% of subjects selected the gray disk as the odd color out, which is predicted by neither view and is again consistent with deferring to an "odd shape out" strategy.

For the triplet consisting of {*bluish disk, gray banana, gray disk*}, memory color theory predicts that the objects should appear to be three different colors — the gray banana should appear yellow, the gray disk should appear gray, and the bluish disk should appear blue — and thus that there should be no salient "odd color out"; in that case, any option should perhaps be equally likely as another. By contrast, the modular view straightforwardly predicts that subjects should pick the bluish disk as the odd color out, since the other two objects are gray. In fact, there *was* a salient odd color out: 69.6% of subjects selected the bluish disk as the odd color out (the choice consistent with the modular view), whereas 0 subjects selected the blue banana as the odd color out, and 30.4% of subjects selected the gray banana as the odd color out; $\chi^2(2, N=148)=108.20$, $p<.001$.

For the triplet consisting of {*gray banana, bluish disk, bluish banana*}, memory color theory predicts that the objects should appear to be three different colors — the gray banana should appear yellow, the bluish banana should appear gray, and the bluish disk should appear blue — and thus that there should be no salient "odd color out"; in that case, any option should perhaps be equally likely as another. By contrast, the modular view straightforwardly predicts that subjects should pick the gray banana as the odd color out, since the other two objects are blue. In fact, there *was* a salient odd color out as indicated by subjects' responses: 68.9% of subjects selected the gray banana as the odd color out (the choice consistent with the modular view), whereas 0.7% of subjects selected the blue banana as the odd color out, and 31.4% of subjects selected the bluish banana as the odd color out; $\chi^2(2, N=148)=104.06$, $p<.001$.

In other words, across these four triplets, subjects' responses always favored the modular account and never favored the memory color account, even when memory color theories made clear predictions about which objects should look different than the others in a given triplet.

Indeed, even if the memory color effects in our study somehow varied in strength relative to previous studies (even though we used the same stimuli as Witzel, 2016, under closely matched conditions), it is striking just how unpopular the memory color theory's predicted "odd-one-out" was for subjects. For example, suppose that for the triplet consisting of {*gray disk, bluish banana, bluish disk*} (shown in Figure 4a), it turned out that the memory color effects was somehow weaker than in previous work, such that the bluish banana was perceptually biased only halfway towards achromatic gray, rather than completely towards achromatic gray as in previous studies. Even then, memory color theories should predict that the gray disk and bluish disk should be chosen roughly equally often, and that the bluish banana should be chosen least often (since it would be most perceptually similar to the other two images); but this pattern was not observed either — instead, subjects just behaved as though they saw the relative coloring accurately and without distortion.

Overall, we took these general patterns of results as initial evidence that memory color effects fail to obey the "logic" expected of bona fide perceptual effects.

Experiment 2: "Where were the bananas?"

Experiment 1 suggested that subjects fail to respond according to the memory color theory in cases with clear predictions about which objects should look similar and which should look different. However, one possibility is that our task caused subjects to focus too closely on the particular colors of certain pixels on the display, and perhaps thereby fail to represent the images as *objects* with known colors. This could undermine the validity of the results, since it is critical to memory color effects that subjects represent canonically colored objects *as those objects* — i.e. that they are representing the banana *as a banana* while viewing it.

To give memory color effects the best chance of revealing themselves in this task, Experiment 2 included a secondary task after each odd-color-out judgment, in which the objects disappeared from the screen and subjects had to identify the locations of all the banana images that had been present a moment earlier. Since the images were no longer on the display during this secondary task, accurate performance on it required subjects to have earlier noticed which objects were bananas and which weren't (or, at least, to have encoded their shapes in some way, rather than just the colors of a few pixels). This encouraged subjects to represent the bananas *as bananas* while making color judgments about them, since subjects knew at that time of viewing the stimuli that they would later have to report the locations of the bananas from memory.[4] This experiment also

[4] We thank Molly O'Rourke-Friel for a comment that inspired this design.

served as a replication of Experiment 1 (and otherwise proceeded in exactly the same way), to ensure the reliability of the relevant patterns.

Method

This experiment was identical to Experiment 1 except as follows. A new group of 250 subjects participated. (We conservatively increased the sample size because of an additional exclusion criterion related to the secondary task.) After making each odd-color-out judgment, the images disappeared and were replaced 500ms later by three empty boxes in the same locations as the trial images. Subjects were asked “Where were the bananas?”, and could click as many or as few of the boxes as they liked; when they were satisfied with their answer, subjects clicked a button labeled “I’ve chosen the bananas”, after which the boxes disappeared for 2000ms and were then replaced by the images for the next trial (Figure 5).

To ensure that we only analyzed data from subjects who were representing the bananas *as bananas*, we excluded any subject who failed to perform above 90% accuracy across all of the “Where were the bananas?” trials; this resulted in the additional exclusion of 7 subjects (2.8% of the total sample).

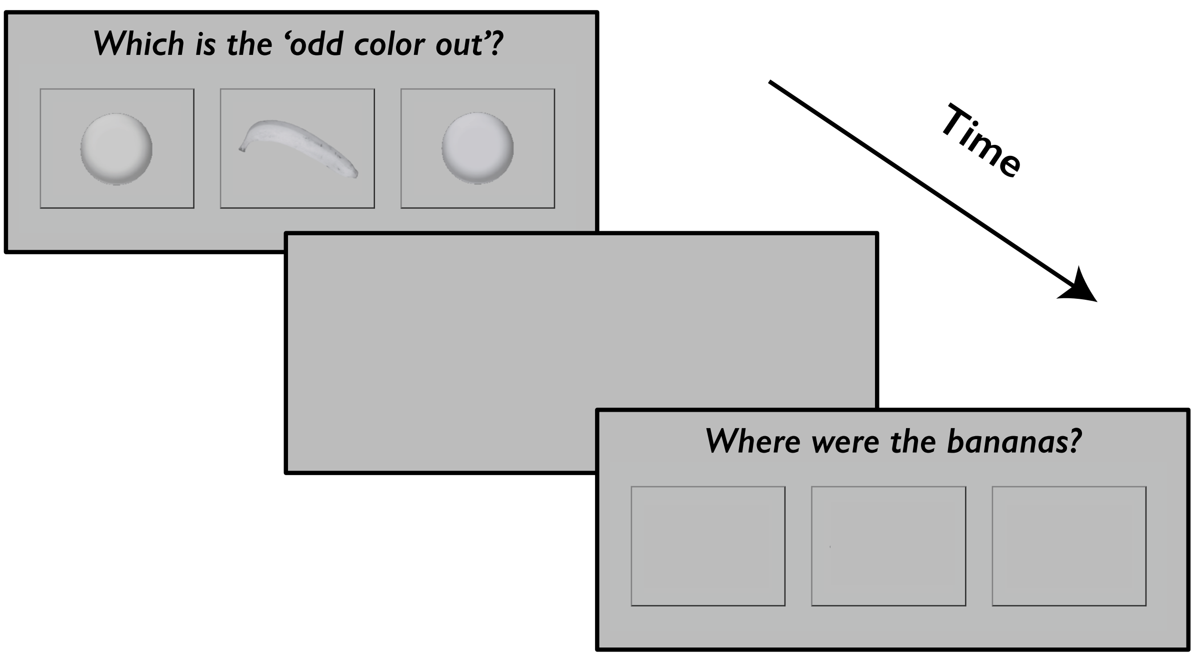

Figure 5. Design of Experiment 2. After picking the odd color out from the display, subjects were asked to recall the locations of the banana(s) that had been on the screen. This required subjects to focus on the identity of the objects while making their initial choice, thereby encouraging subjects to represent the banana images *as bananas*.

Results and Discussion

Every finding from Experiment 1 was replicated in Experiment 2.

For the triplet consisting of {*gray disk, bluish banana, bluish disk*}, 73.9% of subjects selected the gray disk as the odd color out (the choice consistent with the modular view), and only 2.5% of subjects (just two subjects in the entire sample) selected the blue disk as the odd color out (the choice consistent with memory color).

For the triplet consisting of {*bluish banana, gray disk, gray banana*}, 70.8% of subjects selected the bluish banana as the odd color out (the choice consistent with the modular view), and only 5.0% of subjects selected the gray banana as the odd color out (the choice consistent with memory color).

For the triplet consisting of {*bluish disk, gray banana, gray disk*}, 75.6% of subjects selected the blue disk as the odd color out (the choice consistent with the modular view); memory color predicts that all three objects should appear different colors, and thus that any option should be as likely a response as any other.

For the triplet consisting of {*gray banana, bluish disk, bluish banana*}, 73.9% of subjects selected the gray banana as the odd color out (the choice consistent with the modular view); memory color predicts that all three objects should appear different colors, and thus that any option should be as likely a response as any other.

In other words, every triplet showed the pattern predicted by the modular view, and none of them showed the pattern predicted by the memory color view, even among subjects who were actively representing the bananas *as bananas*. Indeed, if anything, these patterns were stronger here than in Experiment 1, despite the increased focus on the bananas' identities.

This result further suggests that memory color effects do not obey the "logic" expected of genuine perceptual effects, and instead behave exactly as they should if they are simply perceived without this sort of distortion. For example, if bluish bananas truly look gray, then they should resemble gray disks (recall that blue shade used here was specifically chosen by memory color researchers to match the shade that subjects choose for a banana to be "gray"); however, we found the opposite pattern — subjects' answers were *least* consistent with the memory color theory, and most consistent with the traditional view that blue objects look blue and gray objects look gray, no matter the subject's knowledge about their canonical colors.

Experiment 3: Perceptual logic in classical memory color effects

The previous experiments explored a new way to study alleged effects of knowledge on perception, by asking whether memory color effects obey a "logic" that should be expected of genuinely perceptual phenomena. In focusing on the strongest and most recent work on memory color effects, however, these studies dealt with only one sort of stimulus, and only one sort of claim. How generally can this strategy be applied?

To answer this question, Experiment 3 turned from contemporary work on memory color effects to the classical investigations that inspired this more recent work — in particular, a report from the middle of the last century that heart-shapes appear redder than identically colored shapes that don't have strong color associations (Delk & Fillenbaum, 1965). In that study, subjects viewed shapes cut out of orange-red cardboard, which appeared against a color-adjustable background. The subjects' task was to adjust the background to match the color of the presented shape (by giving instructions to an experimenter), and the results showed that subjects selected a redder background for the heart than they did for shapes without canonical colors (e.g., circles, triangles, or rectangles).

On one hand, this earlier result may seem weaker than more recent memory color work, in that the study relied on methods that seem especially prone to bias: For example, the experimenters themselves operated the dials that adjusted the background's color, which could have contaminated the results in favor of the experimenters' hypotheses (Gilder & Heerey, 2018). At the same time, one relative strength of these studies is that they involved *matching* the colors of two stimuli, rather than adjusting one stimulus to some internal standard (cf. the achromatic adjustment method of Hansen et al., 2006). As noted earlier, relying on the subject's own notions of such color categories can pose problems for isolating perceptual effects per se, given the contextual flexibility of such subjectively defined color standards. Perceptual matching tasks such as these may also be less susceptible to alternative explanations based on differences in memory rather than perception (Cooper et al., 2012; Firestone & Scholl, 2015b), since they involve affirming the similarity of two currently visible stimuli. In light of these factors, and in light also of this classic study's prominence in contemporary debates over cognitive (im)penetrability (Brogaard & Gatzia, 2017; Deroy, 2013; Gatzia, 2017; Gross et al., 2014; MacPherson, 2012; Stokes, in press; Vetter & Newen, 2014; Zeimbekis, 2013), we asked whether this phenomenon might also be susceptible to a test of perceptual "logic".

El Greco, juiced-up

Our study in this vein is a variant on the "El Greco fallacy" — an episode from art history that has also become a technique for separating perception from judgment (Firestone, 2013b; Firestone & Scholl, 2014; Martin et al., 2016). El Greco famously painted figures that were unusually elongated, and it was once theorized that this reflected a distortion in the Spanish renaissance artist's vision due to unusually severe astigmatism, which was said to vertically blur his perception of the world. However, if El Greco truly experienced a vertically stretched-out world, then he would also have experienced a vertically stretched-out *canvas*, and the distortions would have 'canceled out'. So, whether or not El Greco had astigmatism, that couldn't explain the distortions in his paintings. For perception research, the moral is the same: If an alleged effect is truly perceptual, and if the 'equipment' used to measure this effect it itself similarly susceptible to the manipulation, then the effect should disappear when the manipulation is applied to both the stimulus and the measuring equipment.

Here, we develop an even stronger and more comprehensive version of the El Greco fallacy than has been used in previous research, to ask whether classical memory color effects obey the "logic" of perception: We run *four* conditions of the original Delk & Fillenbaum study, ratherthan two, corresponding to all possible pairs of backgrounds and foregrounds made up of hearts and rectangles. In other words, subjects not only saw hearts and rectangles on rectangle-shaped backgrounds (as in the original study), but also hearts and rectangles on *heart*-shaped backgrounds[5] (Figure 6).

[5] We thank Eli Shupe for a comment that inspired this design. Gross et al. (2014) also take a similar approach.

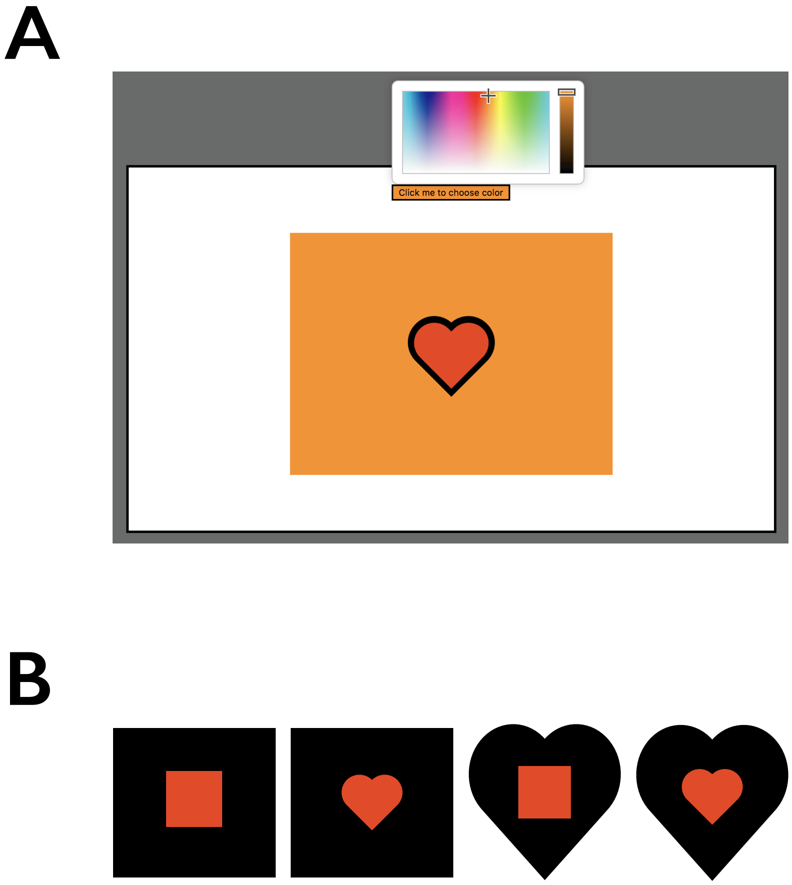

Figure 6. Design of Experiment 3. (A) Subjects saw a shape on a color-adjustable background, and estimated its color by continuously adjusting the color of the background shape to match the color of the foreground shape. (B) The four trial types included a rectangle on a rectangular background, a heart on a rectangular background, a rectangle on a heart-shaped background, and a heart on a heart-shaped background.

The shape of the background itself is relevant because, if hearts look redder than rectangles, then heart-shaped *backgrounds* should themselves appear redder than rectangle-shaped backgrounds, and subjects' color matching judgments should incorporate this distortion too. However, while memory color theories make clear and strong predictions about such cases, those predictions can be difficult to quickly and reliably wrap one's mind around. For example, try to quickly determine how estimates for an orange-red rectangle on a heart-shaped background should differ from estimates for an orange-red heart on a rectangle-shaped background. Memory color theory is committed to just as strong a prediction about such a case as it is in the original case, but we may find ourselves struggling to immediately and confidently articulate this prediction. (The answer is that a rectangle on a heart-shaped background should produce an objectively more orange estimate than a heart on a rectangle-shaped background, in part owing to the extra redness present in the adjustable background.) Indeed, for each of the pairwise comparisons between these conditions, the memory color theory makes an equally strong prediction; the present experiment exhausts these predictions and asks whether they are confirmed.

Method

Subjects

400 subjects were recruited online from Amazon Mechanical-Turk and were monetarily reimbursed.

Stimuli

The stimuli in this experiment consisted of either a love-heart shape or a rectangle, both appearing in a red-orange fill with a thin black outline. The precise shapes and colors used in the original Delk and Fillenbaum study were either unknown or difficult to reproduce on a computer monitor; however, we chose a conventional love-heart shape for the heart, and we chose an orange-red color for the foreground shapes corresponding to RGB(242,59,13), or its equivalent HSV(12°,95%,95%). This color is naturally judged as an example of "red" — and is often called "orange red" in web color guides — but still leaves room in the red-ward direction of the color space. (In HSV color space, red is conventionally located at 0°, and orange is conventionally located at 30°; our sample was located at 12°.) The background against which these shapes appeared was either a larger rectangle, or a larger version of the same heart shape. As earlier, the nature of online data collection means that we cannot be sure of the exact color, size, or brightness (etc.) of the images as they appeared during the experiment; however, any distortions or miscalibrations caused by a given subject's monitor would have been held constant across conditions (just as in previous work studying memory color effects online; Witzel, 2016).

Procedure

Subjects were instructed to "adjust the color of the background until it looks the same as the color of the object in the middle", and they completed one trial of each of four trial types, corresponding to the four pairs of foreground and background shapes: {*rectangle-background*, *rectangle-foreground*}, {*rectangle-background*, *heart-foreground*}, {*heart-background*, *rectangle-foreground*}, {*heart-background*, *heart-foreground*} (Figure 6). The trials appeared in a newly randomized order for each subject.

On each trial, the background began colored in black, and subjects clicked a button to reveal a color palette through which they could navigate using their cursor; as the cursor moved through the space, the background's color changed to match the cursor's location in the color space. Subjects clicked a button to indicate that they were satisfied with the match, at which point the next trial appeared.

Readers can experience the task for themselves at http://perceptionresearch.org/hearts.

Results and Discussion

Following previous work (Gross et al., 2014), we analyzed responses in terms of the degree-difference in hue within the HSV color space, which allows the analysis to collapse over differences in saturation or brightness and instead isolate the "redness" vs. "orangeness" of responses. The degree-value of a given color response represents its angular position within a cylindrical color space: 0° is red, 60° is yellow, 120° is green, etc.

Given the sensitivity of color matching to extreme values (where a single "random" response by a single subject can throw off dozens or hundreds of subtle responses by other subjects), we excluded any subject whose response on any trial was more than 60° away from the object's true color; this is equivalent to answering that a deeply blue object is pink, or a purely red object is yellow, and so would seem to indicate a lack of engagement or understanding on the part of the subject. We also excluded any subject who failed to contribute a complete dataset. This left 363 subjects of the original 400. (We also used these exact same exclusion criteria in a replication experiment; see Experiment 4.)

The results of all four conditions appear together in Figure 7a, and are plotted as the bias in hue *toward red* from the foreground image's true hue (which was 12° in every condition).[6] Below, we consider the various pairwise comparisons that are possible between these conditions, and whether the memory color prediction was in fact observed.

Replication of Delk and Fillenbaum

We first examined the effect from the original Delk and Fillenbaum study, which had found that hearts are judged as redder than familiar shapes that don't have strong color associations; in our

[6] To determine the mean H value of subjects' color settings, rather than the red-ward bias we report here, you could subtract the values in Figure 7a from 12°; for example, for the {*rectangle-background*, *rectangle-foreground*} condition, the red-ward bias was 1.39°, which means subjects set the background to an average H value of 12° – 1.39° = 10.61°.

experiment, this was equivalent to the {*rectangle-background*, *heart-foreground*} vs. {*rectangle-background*, *rectangle-foreground*} contrast (Figure 7b, comparison i). We successfully replicated this effect: In our sample too, subjects adjusted the background rectangle to be redder when the foreground was a heart than when the foreground was a square: 2.84° red-ward vs. 1.39° red-ward, $t(362)=4.17$, $p<.001$. Though this effect is rather small in terms of raw degrees of hue, this result establishes the reliability of Delk and Fillenbaum's original finding — and indeed this may be the first study in several decades to do so. Color estimates for hearts truly are redder than estimates for identically colored squares.

<u>Exhausting the predictions of memory color theories</u>

Is this effect truly perceptual? Having established the reliability of the key rectangle vs. heart contrast, we can now examine other contrasts about which the memory color theories make equally strong predictions. For example, consider the contrast between {*heart-background*, *heart-foreground*} and {*rectangle-background*, *rectangle-foreground*} (Figure 7b, comparison ii): In both cases, a given foreground shape is matched to an identical background shape, and so there should be no effect of the shapes' identity; if hearts appear redder than rectangles, then both the foreground heart and the background heart should appear redder, and the effects should cancel out, since the mind would also have added some extra redness to the background heart. However, we *did* observe an effect between these two conditions: Subjects judged a heart to be redder than a rectangle even when the background of the heart was itself a heart: 2.41° red-ward vs. 1.39° red-ward, $t(362)=2.82$, $p=.005$. This pattern exemplifies the characteristic "El Greco fallacy" result; if hearts truly look redder, then there should have been no difference between these two cases.

Importantly, however, the design of this experiment permits even more comprehensive and powerful tests of the memory color theory's predictions. Even beyond the "El Greco" pattern, we can consider other contrasts — for example, the contrast between {*rectangle-background*, *rectangle-foreground*} and {*heart-background*, *rectangle-foreground*} (Figure 7b, comparison iii). Here, with the foreground shape held constant, subjects should adjust the heart-shaped background to be more *orange* (i.e., less red) than the rectangle-shaped background, to account for the added redness that the mind allegedly adds to hearts. However, we did not observe this effect, and indeed if anything we observed the *opposite* effect: Subjects adjusted the background to be *redder* in the {*heart-background*, *rectangle-foreground*} condition than in the {*rectangle-background*, *rectangle-foreground*} condition: 2.32° red-ward vs. 1.39° red-ward, $t(362)=2.12$, $p=.03$ — the *reverse* of the memory color prediction. (Note that it is not particularly crucial that this "opposite" effect be statistically significant; the key result is simply that it fails to differ in the other direction.) This result is perhaps even more powerful evidence against a perceptual interpretation than the canonical "El Greco"-style result, because it is a minimal pair with the original Delk and Fillenbaum (1965) result: Switching the heart from the foreground to the background should produce the opposite of the original effect, but it does not.

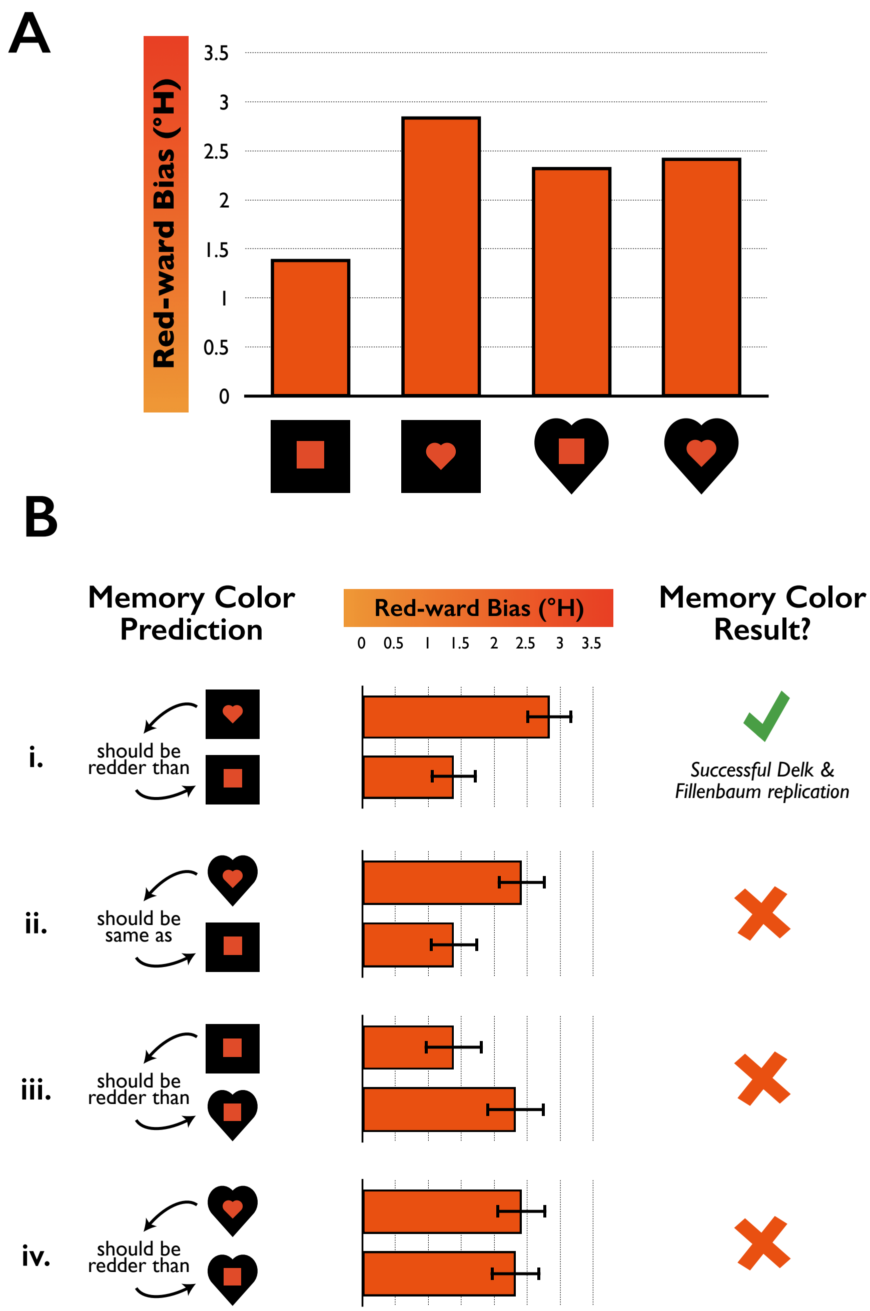

Figure 7. Results of Experiment 3. Whereas a simple contrast of heart vs. rectangle (on a rectangular background) produced the effect expected by memory color theories, other contrasts failed to produce the effects predicted by memory color theories, or even produced *opposite* effects. Error bars for each contrast are ± 1 SE of the difference between conditions.

Consider further the contrast between {*heart-background*, *heart-foreground*} and {*heart-background*, *rectangle-foreground*} (Figure 7b, comparison iv); this contrast should behave exactly like the original Delk and Fillenbaum contrast of {*rectangle-background*, *rectangle-foreground*} vs. {*rectangle-background*, *heart-foreground*}, with a redder estimate for the foreground heart than for the foreground rectangle — since the background is held constant across both conditions, and only the foreground shape has changed. However, there was no effect in this case — 2.41° red-ward vs. 2.32° red-ward, $t(362)=0.24$, $p>.80$ — even though we had indeed observed a robust effect in the {*rectangle-background*, *rectangle-foreground*} vs. {*rectangle-background*, *heart-foreground*} case.

We can also consider {*rectangle-background*, *heart-foreground*} vs. {*heart-background*, *heart-foreground*}; here, {*rectangle-background*, *heart-foreground*} should produce a redder estimate (since the foreground is constant but the rectangular background has no added redness), but in fact there was no effect: 2.84° red-ward vs. 2.41° red-ward, $t(362)=1.25$, $p=.21$. (However, we note that this non-significant trend was indeed in the direction predicted by memory color theory, and it could emerge with greater statistical power; for this reason, we do not wish to read too much into this non-effect here.)

Finally, {*rectangle-background*, *heart-foreground*} vs. {*heart-background*, *rectangle-foreground*} should have produced the largest effect of all, with {*rectangle-background*, *heart-foreground*} producing the reddest estimate of all conditions and {*heart-background*, *rectangle-foreground*} producing the least red estimate of all conditions; however, there was no effect here too: 2.84° red-ward vs. 2.32° red-ward, $t(362)=1.23$, $p=.22$. Although this trend was in the direction predicted by memory color theory, we note that it was numerically *smaller* than the effect it was supposed to be significantly *larger* than: The difference between {*rectangle-background*, *rectangle-foreground*} and {*heart-background*, *rectangle-foreground*} was 1.45°, and so the difference between {*rectangle-background*, *heart-foreground*} vs. {*heart-background*, *rectangle-foreground*} should have been even larger than that difference, according to memory color theories; but instead the {*rectangle-background*, *heart-foreground*} vs. {*heart-background*, *rectangle-foreground*} difference was only 0.51°. In other words, even if this 0.51° effect were statistically significant in a larger sample, it is still the wrong kind of effect to vindicate the memory color theory's prediction.

Thus, whereas we successfully replicate the original heart vs. rectangle effect on a rectangle-shaped background — exactly the kind of hypothesis that might be easier for subjects to follow — we repeatedly fail to confirm the memory color prediction for other cases, or we even actively find the *opposite* pattern. Indeed, just a single failed prediction among the set above would be enough to frustrate the memory color account; however, we found that the memory color theory's prediction was wrong about a majority of its predictions, just as would be expected if the relevant effects were not perceptual.

Alternative explanations?

One possible alternative explanation for some of the above results is that we've mischaracterized the role of the added redness of the background heart. In particular, perhaps the added redness of a heart-shaped background *subsumes* whatever shape appears inside of it, making it appear redder as well (as if the foreground were "inside" the heart-shaped background, taking on the heart's reddish glow). This could, perhaps, explain why heart-shaped backgrounds don't drive judgments orange-ward, since they also imbue their foreground shapes with an equal amount of added redness.

However, this alternative is contradicted by other results: For example, if the effect of a heart-shaped background is additive with whatever is in the foreground, then a heart on a heart-shaped background should have been judged as redder than a rectangle on a heart-shaped background; but this was not the case (indeed, these conditions were the most similar of any condition we tested). And if the effect is not additive — i.e. if the foreground shape gets as much extra redness as it would ever get as long as there is one heart in the foreground or the background — then this alternative fails to explain why a heart on a rectangle-shaped background produced a response no different than a heart on a heart-shaped background. Moreover, both such accounts fail to explain the *redder* judgment in {*heart-background*, *rectangle-foreground*} than in {*rectangle-background*, *rectangle-foreground*}.

Another possibility is that background hearts are somehow not recognized as hearts, or are not attended to as strongly, in such a way that memory color effects don't apply to them. Could this explain our results? First, we note that this account seems unmotivated from the perspective of the memory color framework; to our knowledge, there has not been any prior evidence or suggestion that attention was required for memory color effects. Indeed, memory colors are meant to assist color processing across a whole scene; so if it turned out that memory color effects were hyper-local in this way — applying to one small region of an image but not to the immediately surrounding region — then memory color effects would not be nearly as useful for perception as they are meant to be. But second, this account would fail to explain our results even if it *were* the case that memory color effects don't apply to the background shapes. For example, it would fail to explain why there was a large square vs. heart difference when the background was rectangular (Figure 7b, comparison i), but no square vs. heart difference when the background was heart-shaped (Figure 7b, comparison iv); if the background is ignored for the purposes of memory color effects, then those two contrasts should have produced similar results.

Overall, we thus took this general pattern of results as promising evidence against a perceptual interpretation of classical memory color effects, which seem not to behave as they should if they were genuinely perceptual.

Experiment 4: Replication

Experiment 3 produced results that were inconsistent with what a memory color account would predict; but what explains those results? Examining the totality of the data collected (e.g., Figure 7a), it is difficult to find a single unifying explanation. However, one plausible account that seems to qualitatively cover the relative judgments made in the various conditions is a simple estimation strategy that might be summarized as follows: "*If there is a heart on the display, give a redder estimate*". This simple rule, if implemented by subjects, could account for the fact that every condition involving a heart produced a response that was redder than the baseline {*rectangle-background*, *rectangle-foreground*} condition. And it could also explain why the more fine-grained predictions made by memory color theories failed to come true: The subjects simply weren't making the sophisticated and often tortuous inferences implied by the theory. However, whereas we had actively predicted that many of the conditions from Experiment 3 would disconfirm the predictions made by memory color theory, our present "if it has a heart, answer red" hypothesis was generated post-hoc, having occurred to us only after looking at the data.

Thus, to further support this interpretation, Experiment 4 replicated Experiment 3, but asked each subject to contribute more estimation data, and also debriefed subjects about the purpose of the experiment; this allowed us both to determine the reliability of the pattern observed in Experiment 3 and also to gain further insight into the thought process of subjects completing the experiment.

Method

All methods in Experiment 4 were identical to Experiment 3 except as noted here. 500 subjects were recruited online from Amazon Mechanical-Turk and were monetarily reimbursed. (We conservatively increased the sample size here because we expected to exclude more subjects; see below.) Rather than complete only one trial of each trial type, subjects completed 10 trials of each trial type; the trials appeared in blocks of four (one for each of the four trial types), with the trial order randomized within each block.

Given the role of demand characteristics in producing related sorts of effects (Durgin et al., 2009), subjects were also debriefed about the purpose of the experiment once they had completed all the trials. In particular, they were asked yes-or-no questions about various hypotheses, after the open-ended question, "What did you think was the purpose of this study? Please answer in two sentences".

Results and Discussion

Every result from Experiment 3 replicated in Experiment 4.

We applied the same exclusion criteria as in Experiment 3, excluding any subject who ever gave a response whose hue was more than 60° off the foreground image's true color, or who failed to contribute a complete dataset; since each subject completed 40 trials instead of 4, this inevitably resulted in a higher exclusion rate, leaving 398 subjects of the original 500.

As in Experiment 3, {*rectangle-background*, *heart-foreground*} vs. {*rectangle-background*, *rectangle-foreground*} replicated Delk and Fillenbaum (1965): $t(397)=2.63$, $p<.01$. {*heart-background*, *heart-foreground*} vs. {*rectangle-background*, *rectangle-foreground*} produced an "El Greco" fallacy: $t(397)=2.21$, $p=.028$. {*rectangle-background*, *rectangle-foreground*} vs. {*heart-background*, *rectangle-foreground*} failed to produce an effect in the direction predicted by memory color theories, and if anything again produced an effect in the "wrong" direction: $t(397)=2.30$, $p=.022$. {*heart-background*, *heart-background*} vs. {*heart-background*, *rectangle-foreground*} showed no effect where there "should" have been one: $t(397)=0.09$, $p>.90$. {*rectangle-background*, *heart-foreground*} vs. {*heart-background*, *heart-foreground*} also produced no effect where there "should" have been one: $t(397)=0.24$, $p>.80$). Finally, {*rectangle-background*, *heart-foreground*} vs. {*heart-background*, *rectangle-foreground*} showed no effect where there "should" have been the *largest* effect: $t(397) = 0.15$, $p>0.85$. These results, though overall weaker in magnitude, confirmed the pattern of results from Experiment 3.

Subjects' hypotheses

The present results were also consistent with an estimation strategy that connects the presence of a heart with redder estimates. In Experiment 4, only the condition without a heart in it

— {*rectangle-background*, *rectangle-foreground*} — had an estimate that was less red than any other condition (though, again, other comparisons could reveal reliable alternative effects with larger samples).

This account is also consistent with responses given by subjects when asked about the purpose of the study. Though we do not attempt a systematic coding and analysis of these open-ended responses here, we note anecdotally that many subjects explicitly articulated Delk and Fillenbaum's original hypothesis, with striking clarity, when simply asked in an open-ended way what they thought the experiment was testing. For example:

"maybe people tend to put the heart a little more red"
"To see if the shape of the object changed our perceptions of the color of it"
"If you associated red with the heart shaped item, even if the color was more orange"
"To see how peoples perceptions of color changes with shapes. Maybe people see hearts as a darker red."
"I think this study was about how shapes affect color perception."
"To see if people make the heart shapes more red colored even if they are more orange in hue."
"If the shape affected the color choice. Like maybe I see hearts as more red."
"It was maybe about shape and how we perceive its color. For example we usually instinctively think red when we see a heart."
"If it's a heart you're more likely to choose a more red color. If it's a square, you'd pick more orange."
"Maybe to see if the shape affected the color choice"
"I think the purpose of the study was to see whether shapes affect color perception, maybe"
"You probably were looking to see a relationship between color matching and the type of shape."
"to see if i would rate hearts as more red in color"

Taken together, these results imply that memory color effects of this sort may not reflect changes in visual appearance: Not only do these effects fail to obey various "logical" constraints, but there are available explanations of the effects in terms of strategic or compliant responding by subjects.

General Discussion

Does knowing an object's typical color change its color appearance? Whereas a long-standing research tradition suggests that it does, we extended such claims to new scenarios and circumstances where the underlying theories make strong and specific predictions that are nevertheless tortuous and difficult to grasp. Across new experiments spanning classical and contemporary work, we found that such scenarios fail to produce the effects expected by memory color theories — and often produce the *opposite* effects. Instead, all such results implied that subjects simply saw the objects' colors in a manner undistorted by their beliefs or prior knowledge, and that any distorted responses that did arise could be readily explained by strategic or compliant responding.

These results may bear on discussions in unusually diverse fields. In vision science, memory color effects have been studied not only as a phenomena unto themselves, but also contributors to color constancy and other core processes of color perception (Witzel & Hansen, 2015; Witzel & Gegenfurtner, 2018). Although our results do not entail — and we do not argue — that color appearance *cannot* be affected by this sort of color knowledge, we contend that pre-existing data, tasks, and stimuli fail to settle the issue, such that more work would be required to show that color knowledge plays this kind of active role in color perception. (For different notions of how higher-level cognition might interact with color perception, see Webster & Kay, 2012, and Winawer et al., 2007, both of whom also use sets of colored objects but in rather different designs.) Beyond this, though, these findings reach further into cognitive science more generally, where memory color effects have been near the center of broader disputes over the cognitive (im)penetrability of perception (Lammers, de Haan, & Pinto, 2017; Lupyan, 2015a; Vetter & Newen, 2014) — and even to philosophy, where the influence of higher-level cognition on color perception is discussed not only with respect to the relationship between cognition and perception (Deroy, 2013; Gatzia, 2017; Macpherson, 2012; Zeimbekis, 2013) but also the rational formation of perceptual beliefs (Siegel, 2012), and even the nature of aesthetic experience (Stokes, 2014). Indeed, philosophical discussions of memory color effects are often particularly concerned, as we are here, with the question of whether memory color effects occur at the level of perceptual phenomenology per se; our results address this question directly and suggest reasons not to accept such claims given the available evidence.

Generally applicable

Beyond the particular implications of the present results, we note further that the experimental design strategy employed here is perfectly general, and could be applied to many other alleged cases of top-down effects on visual appearance. Just as a yellow-distorted blue object should look more similar to gray objects than to blue objects, an object that allegedly appears darker, larger, or closer due to cognitive factors should also show the same pattern.

For example, if positive words truly appear bright and negative words truly appear dark (Meier et al., 2007), then perhaps a positive word presented next to both (a) a neutral word of the same brightness, and (b) a neutral word that is objectively brighter, should appear to resemble the brighter word, and thus the equally bright neutral word should stand out as differently bright. If dartboards truly look larger after a subject hits them with darts (Cañal-Bruland et al., 2010), then a recently hit dartboard presented next to both (a) an un-hit dartboard of the same size, and (b) an un-hit dartboard that is objectively larger, should appear to resemble the larger dartboard object in size, such that the equally sized un-hit dartboard should stand out as different in size. If objects truly look closer when they are desired (Balcetis & Dunning, 2010), then a desired object presented near (a) a neutral object of the same distance, and (b) a neutral object that is objectively closer, should resemble the closer object in distance, and the equally distant neutral object should stand out as different in distance (so too with other reported spatial distortions; Caparos et al., 2015; Fini, Brass, & Committeri, 2015; Harber et al., 2011). Any of these findings would strengthen the case for those alleged top-down effects of cognition on perception (though there may of course be other pitfalls lurking in the background; Firestone, 2013a; Firestone & Scholl, 2016).

Indeed, this strategy is applicable even *outside* questions of how cognition does or does not affect perception: As exemplified by the color illusion in Figure 2, just about any effect on appearance can be studied using "odd one out" tasks of this sort.

Generally ***interpretable***

Another notable property of the present strategy — and in particular of the "odd one out" task explored in Experiments 1-2 — is that the data it yields are *interpretable* in an unusually broad and powerful way. One unfortunate feature of many investigations of top-down effects on perception — and indeed of many studies in psychology as a whole — is that statistically significant "positive" effects are typically easier to interpret than "null" effects. For example, if a study finds that wearing a heavy backpack makes hills look steeper (a la Bhalla & Proffitt, 1999), one might tentatively conclude that a real effect is present; but if you *don't* find this result, you might attribute such a failure to other factors, such as a lack of statistical power or a failure to properly apply the manipulation. Indeed, this issue can even afflict strategies such as the "El Greco fallacy" (Firestone & Scholl, 2014) to at least some degree: If you *find* an effect in an "El Greco" condition, you may infer that a given effect may not be perceptual; but if you don't obtain that result, some other factor could again be the culprit (including even statistical power, as perhaps was the case in Gross et al., 2014, who report a study similar in spirit to our Experiment 3 but who fail to replicate the general Delk and Fillenbaum result in a sample of 25 subjects).

By contrast, the data from the "odd-color-out" task used here are interpretable on *multiple* outcomes, since they can produce statistically reliable results both for or against the relevant theory. For example, consider the {*gray disk, bluish banana, bluish disk*} triplet from Experiment 1, where memory color theory predicts that the blue banana should appear gray (and thus that subjects should pick the bluish disk as the odd color out), and where a modular view predicts that the bluish banana should look similar to the equally blue disk (and so subjects should pick the gray disk as the odd color out). *Either* of these two contrary predictions could be positively supported by a statistically reliable preference to choose one object over another: If subjects pick the blue disk as the odd color out, then that result actively supports memory color theory; but if subjects pick the gray disk as the odd color out (as occurred in our studies), then that result actively opposes memory color theory and supports the modular view. (Perhaps the only hard-to-interpret result would be completely noisy or random responding.) This is a relative strength of this strategy over previous investigations of perception vs. judgment (including Firestone & Scholl, 2014, 2015c), and it makes "odd one out" tasks a promising strategy for the durable challenge of separating perception from post-perceptual judgments and responses (Goldstone et al., 2015; Witt et al., 2015), including even in domains that go beyond questions of modularity and cognitive impenetrability.

Conclusions

How can we separate what we *see* from what we *judge*? Though experimentally distinguishing perception from judgment is often difficult, here we have explored one such way, relying on the "logic" that is obeyed by perception but followed only inconsistently or unreliably by higher-level reasoning. This strategy takes seriously the underlying claims of the relevant theories, and simply exhausts their predictions by testing them in more varied scenarios. We suggest that these results not only recast extant claims about how knowledge does or does not affect perception (here, of color), but also point toward a new and broadly applicable strategy for investigating visual appearance.

Acknowledgements

For comments on previous drafts, we thank Bart Anderson, Jacob Beck, David Delijani, Steven Gross, and Maria Olkkonen; for discussion, we thank members of seminars and reading groups at Johns Hopkins, UMass-Amherst, UNSW, Yale, and York.

Data Availability

All data and materials supporting the experiments in this paper are available at https://osf.io/9tebu/.

References

Adams, G. K. (1923). An experimental study of memory color and related phenomena. *The American Journal of Psychology*, *34*, 359–407.

Adams, W. J., Kerrigan, I. S., & Graf, E. W. (2016). Touch influences perceived gloss. *Scientific Reports*, *6*, 21866.

Adeyefa-Olasupo, I., & Flombaum, J. I. (2018). Tutorial review: Evidence for the memory color effect, 1923–2016. *PsyArXiv*.

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science*, *21*, 147–152.

Banerjee, P., Chatterjee, P., & Sinha, J. (2012). Is it light or dark? Recalling moral behavior changes perception of brightness. *Psychological Science*, *23*, 407–409.

Bannert, M. M., & Bartels, A. (2013). Decoding the yellow of a gray banana. *Current Biology*, *23*, 2268–2272.

Bar, M. (2004). Visual objects in context. *Nature Reviews. Neuroscience*, *5*, 617–629.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Bhalla, M., & Proffitt, D. R. (1999). Visual-motor recalibration in geographical slant perception. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1076–1096.

Block, N. (2016). Tweaking the concepts of perception and cognition. *Behavioral and Brain Sciences*, *39,* e232.

Brogaard, B., & Gatzia, D. E. (2017). Is color experience cognitively penetrable? *Topics in Cognitive Science*, *9*, 193–214.

Bruner, J. S., Postman, L., & Rodrigues, J. (1951). Expectation and the perception of color. *The American Journal of Psychology*, *64*, 216–227.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, *64*, 123–152.

Cañal-Bruland, R., Pijpers, J. R. R., & Oudejans, R. R. D. (2010). The influence of anxiety on action-specific perception. *Anxiety, Stress & Coping*, *23*, 353–361.

Caparos, S., Fortier-St-Pierre, S., Gosselin, J., Blanchette, I., & Brisson, B. (2015). The tree to the left, the forest to the right: Political attitude and perceptual bias. *Cognition*, *134*, 155–164.

Cooper, A. D., Sterling, C. P., Bacon, M. P., & Bridgeman, B. (2012). Does action affect perception or memory? *Vision Research*, *62*, 235–240.

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

Delk, J. L., & Fillenbaum, S. (1965). Differences in perceived color as a function of characteristic color. *The American Journal of Psychology*, *78*, 290–293.

Deroy, O. (2013). Object-sensitivity versus cognitive penetrability of perception. *Philosophical Studies*, *162*, 87–107.

Duguid, M. M., & Goncalo, J. A. (2012). Living large: The powerful overestimate their own height. *Psychological Science*, *23*, 36–40.

Dunning, D., & Balcetis, E. (2013). Wishful seeing: How preferences shape visual perception. *Current Directions in Psychological Science*, *22*, 33–37.

Durgin, F. H., Baird, J. A., Greenburg, M., Russell, R., Shaughnessy, K., & Waymouth, S. (2009). Who is being deceived? The experimental demands of wearing a backpack. *Psychonomic Bulletin & Review*, *16*, 964–969.

Durgin, F. H., DeWald, D., Lechich, S., Li, Z., & Ontiveros, Z. (2011). Action and motivation: Measuring perception or strategies? *Psychonomic Bulletin & Review*, *18*, 1077–1082.

Fini, C., Brass, M., & Committeri, G. (2015). Social scaling of extrapersonal space: Target objects are judged as closer when the reference frame is a human agent with available movement potentialities. *Cognition*, *134*, 50–56.

Firestone, C. (2013a). How "paternalistic" is spatial perception? Why wearing a heavy backpack doesn't—and *couldn't*—make hills look steeper. *Perspectives on Psychological Science*, *8*, 455–473.

Firestone, C. (2013b). On the origin and status of the "El Greco fallacy." *Perception*, *42*, 672–674.

Firestone, C., & Scholl, B. J. (2014). "Top-down" effects where none should be found: The El Greco fallacy in perception research. *Psychological Science*, *25*, 38–46.

Firestone, C., & Scholl, B. J. (2015a). Can you experience "top-down" effects on perception? The case of race categories and perceived lightness. *Psychonomic Bulletin & Review*, *22*, 694–700.

Firestone, C., & Scholl, B. J. (2015b). Enhanced visual awareness for morality and pajamas? Perception vs. memory in "top-down" effects. *Cognition*, *136*, 409–416.

Firestone, C., & Scholl, B. J. (2015c). When do ratings implicate perception versus judgment? The "overgeneralization test" for top-down effects. *Visual Cognition*, *23*, 1217–1226.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, e229, 1–77.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*, 432–446.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*, 1367–1368.

Harber, K. D., Yeung, D., & Iacovelli, A. (2011). Psychosocial resources, threat, and the perception of distance and height: Support for the resources and perception model. *Emotion*, *11*, 1080–1090.

Gatzia, D. E. (2017). Cognitive penetration and memory colour effects. *Erkenntnis*, *81*, 1–23.

Gilder, T. S. E., & Heerey, E. A. (2018). The role of experimenter belief in social priming. *Psychological Science*, *29*, 403–417.

Goldstein, E. B., & Brockmole, J. (2016). *Sensation and perception*. Cengage Learning.

Goldstone, R. L., de Leeuw, J. R., & Landy, D. H. (2015). Fitting perception in and to cognition. *Cognition*, *135*, 24–29.

Gross, S., Chaisilprungraung, T., Kaplan, E., Menendez, J. A., & Flombaum, J. (2014). Problems for the purported cognitive penetration of perceptual color experience and Macpherson's proposed mechanism. *Baltic International Yearbook of Cognition, Logic and Communication*, *9*.

Lafer-Sousa, R., Hermann, K. L., & Conway, B. R. (2015). Striking individual differences in color perception uncovered by "the dress" photograph. *Current Biology*, *25*, R545–R546.

Lammers, N. A., de Haan, E. H., & Pinto, Y. (2017). No evidence of narrowly defined cognitive penetrability in unambiguous vision. *Frontiers in Psychology*, *8*, 2042.

Lupyan, G. (2015a). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology*, *6*, 547–569.

Lupyan, G. (2015b). Object knowledge changes visual appearance: Semantic effects on color afterimages. *Acta Psychologica*, *161*, 117–130.

Machery, E. (2015). Cognitive penetrability: A no-progress report. In: *The cognitive penetrability of perception: New philosophical perspectives*, eds. J. Zeimbekis & A. Raftopoulos. Oxford University Press.

Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, *84*, 24–62.

Malcolm, G. L., Nuthmann, A., & Schyns, P. G. (2014). Beyond gist: Strategic and incremental information accumulation for scene categorization. *Psychological Science*, *25*, 1087–1097.

Martin, J.-R., Sackur, J., Anlló, H., Naish, P., & Dienes, Z. (2016). Perceiving time differences when you should not: Applying the El Greco fallacy to hypnotic time distortions. *Frontiers in Psychology*, *7*, 433.

Masters, R., Poolton, J., & van der Kamp, J. (2010). Regard and perceptions of size in soccer: Better is bigger. *Perception*, *39*, 1290–1295.

Meier, B. P., Robinson, M. D., Crawford, L. E., & Ahlvers, W. J. (2007). When "light" and "dark" thoughts become light and dark responses: Affect biases brightness judgments. *Emotion*, *7*, 366–376.

Newen, A., & Vetter, P. (2017). Why cognitive penetration of our perceptual experience is still the most plausible account. *Consciousness and Cognition*, *47*, 26–37.

Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2008). Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *Journal of Vision*, *8*:13, 1–16.

Pearce, B., Crichton, S., Mackiewicz, M., Finlayson, G. D., & Hurlbert, A. (2014). Chromatic illumination discrimination ability reveals that human colour constancy is optimised for Blue Daylight Illuminations. *PLoS One*, *9*, e87989.

Proffitt, D. R. (2006). Embodied perception and the economy of action. *Perspectives on Psychological Science*, *1*, 110–122.

Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*, 341–365.d

Robilotto, R., & Zaidi, Q. (2004). Limits of lightness identification for real objects under natural viewing conditions. *Journal of Vision*, *4*, 779–797.

Rock, I. (1983). *The logic of perception*. MIT Press.

Siegel, S. (2012). Cognitive penetrability and perceptual justification, *46*, 201–222.

Stokes, D. (in press). Cognitive penetration and the perception of colour. In: R*outledge Handbook of Philosophy of Colour*, eds. D. Brown & F. Macpherson. Routledge.

Stokes, D. (2014). Cognitive penetration and the perception of art. *Dialectica*, *68*, 1–34.

Schwartz, B. L., & Krantz, J. H. (2017). *Sensation and Perception*. Sage Publications.

van Ulzen, N. R., Semin, G. R., Oudejans, R. R. D., & Beek, P. J. (2008). Affective stimulus properties influence size perception and the Ebbinghaus illusion. *Psychological Research*, *72*, 304–310.

Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, *27*, 62–75.

Webster, M. A., & Kay, P. (2012). Color categories and color appearance. *Cognition*, *122*, 375–392.

White, C. W., & Montgomery, D. A. (1976). Memory colors in afterimages: A bicentennial demonstration. *Perception & Psychophysics*, *19*, 371–374.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*, 7780–7785.

Witt, J. K. (2011). Action's effect on perception. *Current Directions in Psychological Science*, *20*, 201–206.

Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. *Perception*, *44*, 289–300.

Witzel, C. (2016). An easy way to show memory color effects. *i-Perception*, *7*.

Witzel, C., & Gegenfurtner, K. R. (2018). Color perception: Objects, constancy, and categories. *Annual Review of Vision Science*, *4*, 475–499.

Witzel, C., & Hansen, T. (2015). Memory effects on colour perception. In: *Handbook of Color Psychology,* eds. A. J. Elliot, M. D. Fairchild, & A. Franklin. Cambridge, UK: Cambridge University Press.

Witzel, C., O'Regan, J. K., & Hansmann-Roth, S. (2017). The dress and individual differences in the perception of surface properties. *Vision Research*, *141*, 76–94.

Witzel, C., Olkkonen, M., & Gegenfurtner, K. R. (2018). A Bayesian model of the memory colour effect. *i-Perception*, *9*, 1–16.

Witzel, C., Valkova, H., Hansen, T., & Gegenfurtner, K. R. (2011). Object knowledge modulates colour appearance. *i-Perception*, *2*, 13–49.

Zeimbekis, J. (2013). Color and cognitive penetrability. *Philosophical Studies*, *165*, 167–175.