#TheDress: Categorical perception of an ambiguous color image

Rosa Lafer-Sousa

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Bevil R. Conway

Laboratory of Sensorimotor Research, National Eye Institute, and National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

We present a full analysis of data from our preliminary report (Lafer-Sousa, Hermann, & Conway, 2015) and test whether #TheDress image is *multistable*. A multistable image must give rise to more than one mutually exclusive percept, typically within single individuals. Clustering algorithms of color-matching data showed that the dress was seen categorically, as white/gold (W/G) or blue/black (B/K), with a blue/brown transition state. Multinomial regression predicted categorical labels. Consistent with our prior hypothesis, W/G observers inferred a cool illuminant, whereas B/K observers inferred a warm illuminant; moreover, subjects could use skin color alone to infer the illuminant. The data provide some, albeit weak, support for our hypothesis that day larks see the dress as W/G and night owls see it as B/K. About half of observers who were previously familiar with the image reported switching categories at least once. Switching probability increased with professional art experience. Priming with an image that disambiguated the dress as B/K biased reports toward B/K (priming with W/G had negligible impact); furthermore, knowledge of the dress's true colors and any prior exposure to the image shifted the population toward B/K. These results show that some people have switched their perception of the dress. Finally, consistent with a role of attention and local image statistics in determining how multistable images are seen, we found that observers tended to discount as achromatic the dress component that they did not attend to: B/K reporters focused on a blue region, whereas W/G reporters focused on a golden region.

Introduction

Most visual stimuli are underdetermined: A given pattern of light can be evidence for many different surfaces or objects. Despite being underdetermined, most retinal images are resolved unequivocally. It is not known how the brain resolves such ambiguity, yet this process is fundamental to normal brain function (Brainard et al., 2006; Conway, 2016). Multistable images are useful tools for investigating the underlying neural mechanisms. The two defining properties of multistable stimuli are that they give rise to more than one plausible, stable, percept within single individuals and that the alternative percepts are mutually exclusive (Leopold & Logothetis, 1999; Long & Toppino, 2004; Schwartz, Grimault, Hupe, Moore, & Pressnitzer, 2012; Scocchia, Valsecchi, & Triesch, 2014). Multistable images are similar to binocular rivalrous stimuli, although in binocular rivalry the competition is between two different images rather than alternative interpretations of a single image. Although the first account of binocular rivalry involved color (Dutour, 1760), to date there are no striking examples of multistable color images. Of course, not all colored stimuli are unambiguous: Consider turquoise, which might be called blue or green by different people. Such ambiguous color stimuli typically retain their ambiguity even when labeled categorically, unlike multistable shape images (Klink, van Wezel, & van Ee, 2012). To date, the best example of something approximating a multistable color phenomenon is the colored Mach card, in which the color of a bicolored card folded along the color interface and viewed monocularly can vary depending on whether one perceives the card receding or protruding (Bloj, Kersten, & Hurlbert, 1999). But it is not clear that the color perceptions of the Mach card are categorical. Moreover, the phenomenon is primarily an illusion of 3-D geometry: Without stereopsis, the perspective cues are ambiguous; the way the colors are perceived is contingent on how these cues are resolved.

Citation: Lafer-Sousa, R., & Conway, B. R. (2017). #TheDress: Categorical perception of an ambiguous color image. *Journal of Vision*, *17*(12):25, 1–30, doi:10.1167/17.12.25.

doi: 10.1167/17.12.25 Received April 13, 2017; published October 31, 2017 ISSN 1534-7362

Could #TheDress be an elusive multistable color image? Initial reports on social media raised the possibility that the image was seen in one of two mutually exclusive ways, as white and gold (W/G) or blue and black (B/K). But color-matching data (not color names) reported by Gegenfurtner, Bloj, and Toscani (2015) concluded that there were many different ways in which the dress's colors could be seen. The tentative conclusion was that reports of two categories arose as an artifact of the two-alternative forced-choice question posed by social media ("Do you see the dress as W/G or B/K?"). The implication was that the true population distribution is unimodal, which is inconsistent with the idea that the image is multistable. The Gegenfurtner et al. study measured perceptions of 15 people. It is not known how many subjects would be required to reject the hypothesis that the population distribution is unimodal. We addressed these issues through a full, quantitative analysis of the results that we presented in preliminary form shortly after the image was discovered, in which we argued that the dress was seen categorically (Lafer-Sousa, Hermann, & Conway, 2015). A side goal was to evaluate the extent to which tests conducted online replicate results obtained under laboratory conditions. Many studies of perception and cognition are being conducted through online surveys; it remains unclear whether results obtained in a lab and online are comparable.

Popular accounts suggest that people are fixed by "one-shot learning" in the way they see the dress image (Drissi Daoudi, Doerig, Parkosadze, Kunchulia, & Herzog, 2017). These observations have been taken to imply that the dress is not like a typical multistable image, because it is widely thought that most people experience frequent perceptual reversals of multistable images. But frequent reversals might not be a necessary property of multistability (see Discussion). The perception of multi-stable shape images at any given instant was initially thought to depend only on low-level factors, such as where in the image one looked (Long & Toppino, 2004): since we move our eyes frequently, it was assumed that the perception of a multi-stable shape image would necessarily reverse frequently. It is now recognized that high-level factors, including familiarity with the image, prior knowledge, personality, mood, attention, decision making, and learning, also play a role in how multi-stable shape images are seen (Kosegarten & Kose, 2014; Leopold & Logothetis, 1999; Podvigina & Chernigovskaya, 2015). These factors are often modulated over a longer time frame than eye movements, which could explain why some multi-stable images do not reverse very often. Data in our initial report suggested that some observers experience perceptual reversals of the dress, raising the possibility that the image is not unlike other multistable images. Here we determined the extent to which the individual differences in perception of the dress image are fixed. We characterized the conditions that promote perceptual reversals of the dress, and tested five factors known to influence how multistable images are perceived: prior knowledge about the image (Rock & Mitchener, 1992); exposure to disambiguated versions (Fisher, 1967; Long & Toppino, 2004); low-level stimulus properties (e.g., stimulus size; Chastain & Burnham, 1975); where subjects look (or attend; Ellis & Stark, 1978; Kawabata & Mori, 1992; Kawabata, Yamagami, & Noaki, 1978); and priors encoded in genes or through lifetime experience (Scocchia et al., 2014). These experiments were afforded because we tested people who varied in terms of both prior exposure to the image and knowledge about the color of the dress in the real world.

By examining the factors that influence perception of the dress image, we hoped to shed light on how the brain resolves underdetermined chromatic signals. While low-level sensory mechanisms like adaptation in the retina can account for color constancy under simple viewing conditions (Chichilnisky & Wandell, 1995; D'Zmura & Lennie, 1986; Foster & Nascimento, 1994; Land, 1986; Stiles, 1959; von Kries, 1878; Webster & Mollon, 1995), they fail to explain constancy of natural surfaces (Brainard & Wandell, 1986; Webster & Mollon, 1997) and real scenes (Hedrich, Bloj, & Ruppertsberg, 2009; Khang & Zaidi, 2002; Kraft & Brainard, 1999). We have argued that the competing percepts of the dress are the result of ambiguous lighting information. The colors of the pixels viewed in isolation align with the colors associated with daylight (Brainard & Hurlbert, 2015; Conway, 2015; Lafer-Sousa et al., 2015). The visual system must contend with two plausible interpretations—that the dress is either in cool shadow or in warm light. In our prior report, we tested the idea that illumination assumptions underlie the individual differences in color perception of the dress, by digitally embedding the dress in scenes containing unambiguous illumination cues to either warm or cool illumination. Most observers conformed to a single categorical percept consistent with the illumination cued (Lafer-Sousa et al., 2015). Here we directly tested the hypothesis by analyzing subjects' judgments about the light shining on the dress. We also tested our hypothesis that the way the dress is seen can be explained by one's chronotype: Night owls spend much of their awake time under incandescent light, and we hypothesized they would therefore be more likely to discount the orange component of the dress and see the dress's colors as B/K; day larks spend more of their time under blue daylight, and we surmised that they would discount the blue component of the dress and see the colors as W/G (Rogers, 2015).

Finally, we used the dress image as a tool to examine the role of memory colors in color constancy. The spectral bias of the illuminant could, in theory, be determined by comparing the chromatic signals entering the visual system with the object colors stored in memory. The gamut of human skin occupies a distinctive profile in cone-contrast space that is surprisingly stable across skin types and shifts predictably under varying illuminations (Crichton, Pichat, Mackiewicz, Tian, & Hurlbert, 2012). These statistics, coupled with the fact that skin is viewable in almost every natural glance, make skin a potentially good cue to estimate the spectral bias of the illuminant (Bianco & Schettini, 2012). We used our disambiguation paradigm (Lafer-Sousa et al., 2015), digitally embedding the dress in scenes in which we systematically introduced different cues to the illuminant. As far as we are aware, the results provide the first behavioral evidence that skin color is sufficient to recover information about the illuminant for color constancy.

Methods and materials

Experimental setup

Detailed methods are provided in the supplementary material of our previous report (Lafer-Sousa et al., 2015). Raw materials and sample analyses are provided here: https://github.com/rlaferso/-TheDress. The majority of participants ($N = 2{,}200$) were recruited and tested online through Amazon's Mechanical Turk using a combination of template (Morris Alper's Turk Suite Template Generator 2014, available online at http://mturk.mit.edu/template.php) and custom HTML and JavaScript code. A smaller number of subjects ($N = 53$) were recruited from the Massachusetts Institute of Technology (MIT) and Wellesley College campus through word of mouth and social media, and tested using the M-Turk platform on a calibrated display in the laboratory. We adhered to the policies of the MIT Committee on the Use of Humans as Experimental Subjects in using Amazon's Mechanical Turk for research purposes. Informed consent was obtained from those subjects who performed the study in the laboratory study. Procedures were approved by the institutional review board of Wellesley College. Subjects were between 18 and 69 years of age. To control for subject quality among the Mechanical Turk participants, we required that subjects have Mechanical Turk approval ratings of 95% or higher and have previously completed at least 1,000 human intelligence tasks on Mechanical Turk.

In-laboratory subjects viewed the display at 40 cm. Subjects used a chin rest to control viewing angle and distance. The experiment was performed on a calibrated 21.5-in. iMac computer with a pixel resolution of 1,920 × 1,080 in a windowless room with LED overhead lighting (CIExyY: 0.4814, 0.4290, 4.3587 cd/m^2), measured off the Macbeth color checker's standard white, held at the same location and viewing angle on the monitor at which we presented the dress image). Normal color vision was confirmed with Ishihara plates (Ishihara, 1977).

To ensure that stimuli were the same size across displays for online subjects, we specified the sizes of stimulus images in absolute pixels in the HTML experiment code. There is some variability from display to display in terms of the actual physical size of a pixel. We measured the images on a typical monitor in the laboratory to provide a reasonable estimate for how the pixel values correspond to degrees of visual angle. We estimate that among different displays the variance in actual display size was $\sim\pm10\%$ of the size measured in lab.

Three experiments were conducted: 1, a main experiment; 2, a follow-up experiment to assess the role of image size in determining what colors people report; and 3, an in-laboratory, controlled experiment. Data were pooled from the various experiments depending on the analysis performed.

In each experiment, dress color percepts were queried using two tasks: a free-response color-naming task and a color-matching task. In Experiments 1 and 3 subjects were also asked to report on their impressions of the lighting conditions in the image by providing temperature ratings and verbal descriptors, and to estimate where in the image they thought they spent most of their time looking. In addition, Experiments 1 and 3 queried color percepts for a set of digitally synthesized test stimuli featuring the dress (cropped from the original image) embedded in scenes with unambiguous simulated lighting conditions, and a version of the original image that had been spatially blurred. Experiment 3 included an additional set of synthetic test images not presented in Experiment 1. Each experiment contained questions about subjects' demographics, viewing conditions, and past viewing experiences, which were distributed throughout the experiment (the full questionnaire is reproduced in Lafer-Sousa et al., 2015).

Stimuli

Full-scale stimuli reproductions are provided in the Supplementary Image Appendix; they can also be viewed online at https://youtu.be/U6c4au-Wu-E.

Original dress image (used in Experiments 1, 2, and 3)

The original dress image circulated on the Internet (courtesy of Cecilia Bleasdale; Figure 1). In Experiments 1 and 3, the original dress photograph was

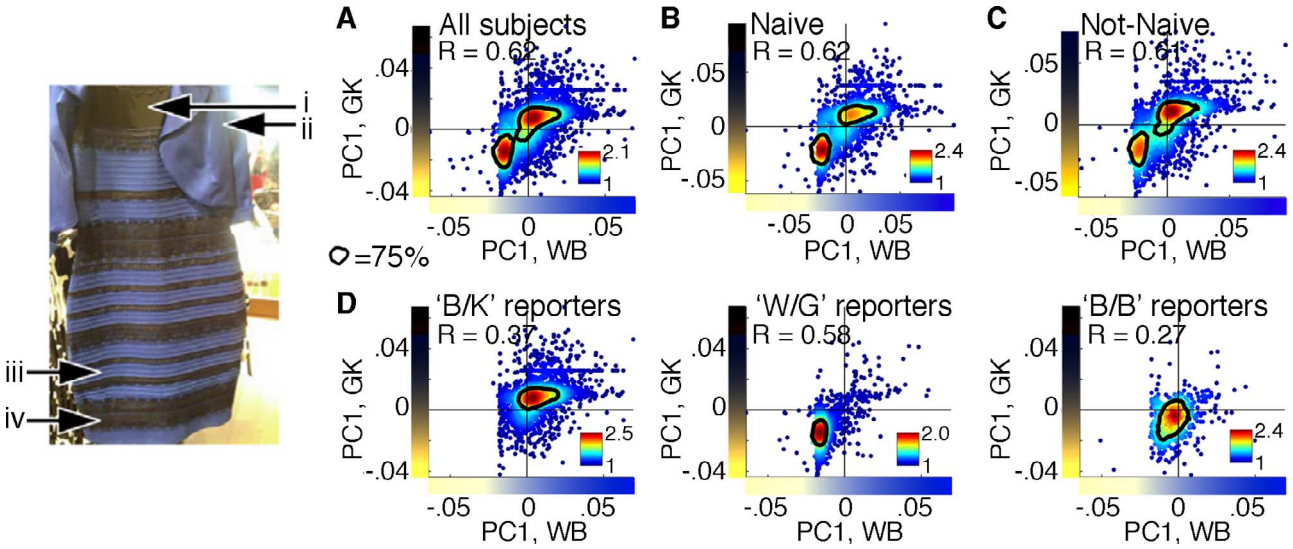

Figure 1. Population distributions of subjects' color matches show categorical perception of the dress. Subjects used a digital color picker to match their perception of four regions of the dress (i, ii, iii, iv); the dress image was shown throughout the color-matching procedure. (A) Matches for regions i and iv of the dress plotted against matches for regions ii and iii, for all online subjects ($N = 2{,}200$; $R = 0.62$, $p < 0.001$). Contours contain the highest density (75%) of matches. The first principal component of the population matches (computed from CIELUV values) to (i, iv) defined the y-axis (gold/black: GK); the first principal component of the population matches to (ii, iii) defined the x-axis (white/blue: WB). Each subject's (x, y) values are the principal-component weights for their matches; each has two (x, y) pairs, corresponding to (i, ii) and (iii, iv). Color scale is number of matches (smoothed). (B) Color matches for regions (i, iii) of the dress plotted against matches for regions (ii, iv) for subjects who had never seen the dress before the experiment (Naïve; $N = 1{,}017$; $R = 0.62$, $p < 0.001$). Axes and contours were defined using data from only those subjects. (C) Color matches for regions (i, iv) of the dress plotted against matches for regions (ii, iii) for subjects who had seen the dress before the experiment ($N = 1{,}183$; $R = 0.61$, $p < 0.001$). Axes and contours were defined using data from only those subjects. (D) Color matches for all subjects (from A) were sorted by subjects' verbal color descriptions ("blue/black" = B/K, $N = 1{,}184$; "white/gold" = W/G, $N = 686$; "blue/brown" = B/B, $N = 272$) and plotted separately. Axes defined as in (A). In all panels, contours contain the highest density (75%) of the matches shown in each plot. Dress image reproduced with permission from Cecilia Bleasdale.

presented at 36% of its original size so that the entire image would be visible on the display. The image was presented at an absolute size of 226 × 337 pixels. This corresponded to 7.2° of visual-angle width. In Experiment 2, the original dress photograph was presented at one of four sizes, defined as a percentage of the original: 10% (63 × 94 pixels, or 2.0° of visual angle on a 21.5-in. iMac), 36% (the image size presented in Experiment 1; 226 × 337 pixels, or 7.2° of visual angle), 100% (628 × 937 pixels, or 20.0° of visual angle), and 150% (942 × 1,405 pixels, or 30.0° of visual angle). The stimuli in the 10% and 36% conditions fit fully in all browser windows. For the 100% and 150% images, only part of the image was visible at once in the height direction, but the full width of the image was completely visible in the horizontal dimension. As a result, subjects had to scroll over the stimulus image, from top to bottom, to view it. Scrolling to see the entire image was required in order for subjects to access the buttons to move through the study, ensuring that all subjects saw the entire image even in the 100% and 150% displays. Subjects were randomly assigned to one of the four scale conditions.

Blurry dress stimulus (used in Experiments 1 and 3)

A blurry version of the original image was presented at 41% of the source size, with a Gaussian blur radius of 3.3 pixels (0.11°). The image was 8.3° of visual angle along the horizontal axis.

Disambiguating stimuli (warm and cool illumination simulations)

Cue-rich test stimuli (used in Experiments 1 and 3)

To create the cue-rich test stimuli (see Supplementary Image Appendix for full-size reproductions), we digitally dressed a White female model in the garment and embedded her in a scene depicting a Rubik's cube under either a simulated warm (yellowish) or cool (bluish) illuminant (cube reproduced with permission from Beau Lotto; Lotto & Purves, 2002). In the cool-illumination scene, the woman was positioned in the shadow cast by the cube. The colors of her skin and hair were tinted to reflect the color bias of the simulated illuminant using a semitransparent color overlay. The chromaticity of the overlay was defined on the basis of

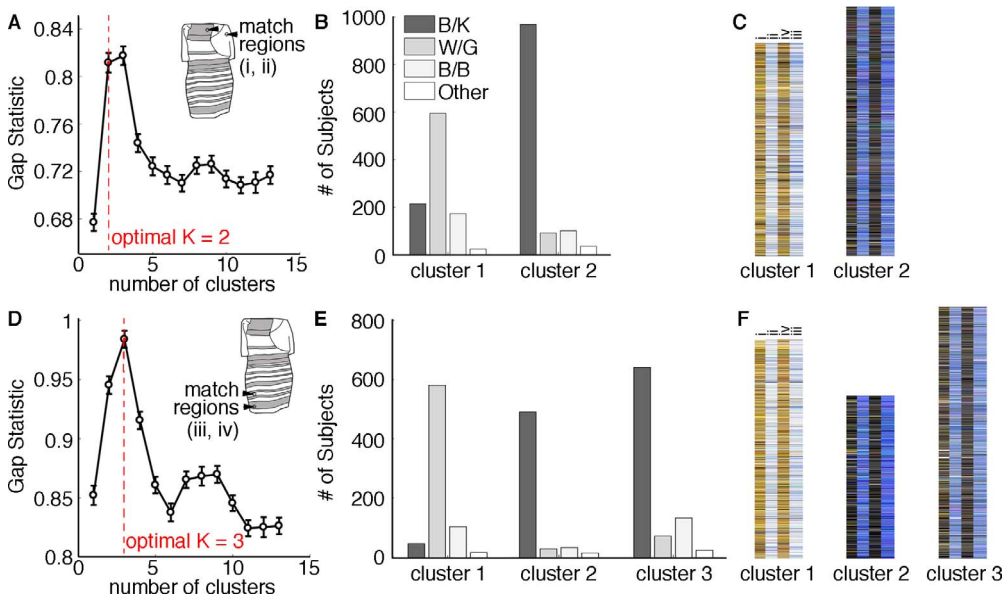

Figure 2. *K*-means clustering of color-matching data favor a two- or three-component model over a single-component distribution. Plots summarize the results from *k*-means clustering assessment (via the gap method; Tibshirani et al., 2001) of the color-matching data presented in Figure 1A ($N = 2{,}200$ subjects). (A) The gap statistic computed as a function of the number of clusters, for color-matching data (principal-component analysis weights) obtained for the upper regions of the dress, using all the data from the online population. Dashed line indicates the optimal *k* clusters. (B) Bar plot showing the cluster assignments of the color matches, binned by the color terms used by the subjects to describe the dress. (C) RGB values of the color matches, sorted by cluster assignment from (B); each thin horizontal bar shows the color matches for a single subject. (D–F) As for (A–C), but for color matches made to the bottom regions of the dress. The gap analysis compares the within-cluster dispersion of a *k*-component model to its expectation under a null model of a single component. The algorithm seeks to identify the smallest number of clusters satisfying $\text{Gap}(k) \geq \text{GAPMAX} - \text{SE(GAPMAX)}$, where *k* is the number of clusters, Gap(*k*) is the gap value for the clustering solution with *k* clusters, GAPMAX is the largest gap value, and SE(GAPMAX) is the standard error corresponding to the largest gap value. The optimal *k* solution for the distribution of upper dress-region color matches is two clusters, and for the lower dress regions it is three, confirming the suspected nonunimodality of the underlying population distribution.

the chromaticity of the white component of the scene's checkered floor, which provides a quantitative white point for the scene (for the cool-illumination scene we used the white checkers that were cast in shadow, corresponding to our placement of the model in shadow). In the warm scene, the white point was 0.352, 0.394, 66 cd/m^2 (CIExyY 1931); in the cool scene it was 0.249, 0.271, 23 cd/m^2. Note that the pixels making up the dress were never manipulated. The dress portion of the stimulus was presented at 76% of the size of the dress in the original image; the complete picture was 518 × 429 pixels (~16.5° of visual angle on the horizontal axis). Throughout this article, figures showing the model are for illustration purposes only: Copyright for the photograph of the model we used to create the stimuli could not be secured for reproduction.

Uniform-surround test stimuli (used in Experiment 3)

To test whether a low-level sensory mechanism like receptor adaptation or local color contrast is sufficient to resolve the dress's colors, we superimposed the isolated dress on uniform fields matched to the mean chromaticity of the cue-rich scenes (CIExyY warm field: 0.363, 0.414, 51 cd/m^2; cool field: 0.276, 0.293, 29 cd/m^2). The dress portion of the stimulus was presented at 76% of the size of the dress in the original image; the complete picture was 518 × 429 pixels (~16.5° of visual angle on the horizontal axis). The pixels that make up the dress were not manipulated.

Skin-only test stimuli (Experiment 3)

To test whether skin chromaticity is by itself a sufficient cue to achieve good color constancy, we presented the dress superimposed on the woman on a white (achromatic) background (CIExyY: 0.322, 0.352, 75 cd/m^2) and tinted her skin according to the spectral bias of the illuminants simulated in the cue-rich scenes. The dress portion of the stimulus was presented at 76% of the size of the dress in the original image; the complete picture was 518 × 429 pixels (~16.5° of visual angle on the horizontal axis). The pixels that make up the dress were not manipulated.

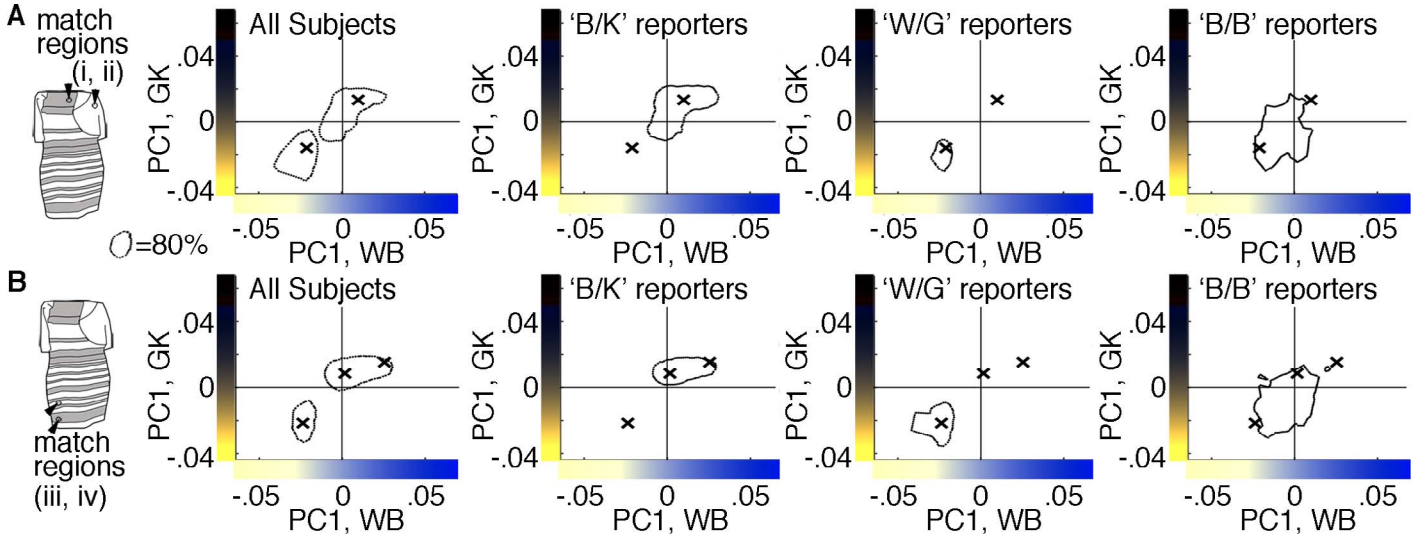

Figure 3. Comparison of color-matching data (contours) with predictions from the *k*-means clustering solutions (x), sorted by subjects' verbal reports. Color-matching distribution contours and *k*-means cluster centroids derived from independent data sets for (A) the color matches made to upper regions of the dress (region i is plotted against region ii; principal-component analysis weights and principal-component axes from Figure 1A) and (B) lower regions of the dress (iv, iii). Distribution contours were determined using one half of the data set (randomly sampled from the online subject pool); cluster centroids were determined using the left-out data (clustered using the optimal *k* identified in the gap analysis from Figure 2). Individual plots show the contours encompassing the top 80% of color matches for half the data from each group (left to right): all subjects; subjects who described the dress as blue and black (B/K); subjects who described the dress as white and gold (W/G); and subjects who described the dress as blue and brown (B/B). Within each row, the same cluster centroids are replotted across the panels, reflecting the outcome of clustering the unsorted data set (i.e., independent of verbal reports).

We considered color constancy good if the majority of subjects conformed to the percept predicted by the lighting cues in each condition, and bad when individual subjects' perceptions were unaffected by the changes in simulated lighting conditions: Under the cool illuminant, good color constancy predicts that subjects should discount a cool (blue) component and see the dress as W/G, while under the warm illuminant they should discount a warm (yellow) component and see the dress as B/K. Note that we use the terms *blue*, *yellow*, *white*, *warm*, and *cool* as shorthand for accurate colorimetric descriptions. McNemar's chi-square tests were used to compare goodness of constancy achieved in different stimulus conditions. McNemar's test is a within-subject z test of equality of proportions for repeated measures. Each test compares the proportion of subjects that did or did not conform to the percept cued in stimulus condition X versus the proportion that did or did not conform to the percept cued in stimulus condition Y. Six tests were performed: cue-rich scene (warm) versus uniform background (warm); skin tint (warm) versus uniform background (warm); skin tint (warm) versus cue-rich scene (warm); cue-rich scene (cool) versus uniform background (cool); skin tint (cool) versus uniform background (cool); and skin tint (cool) versus cue-rich scene (cool).

Tasks

Color naming

Each image was shown for 15 s, and then subjects were prompted to report the apparent color of the dress via a free-response verbal task (two text boxes were provided): "Please look carefully at the dress. A 'continue' button will appear just below this text after 15 seconds. This is a picture of a ____ and ____ dress. (Fill in the colors that you see)." The image was on the screen continuously while the subjects responded (this was not designed as a test of color memory). Color descriptions of the dress were binned into categories: blue/black, white/gold, blue/brown, other (following the methods outlined in Lafer-Sousa et al., 2015).

Lighting judgments

After performing the color-naming task, subjects were prompted to rate the apparent quality of the light illuminating the background of the image ("On a scale from 1 to 5, where 1 is cool and 5 is warm, please rank the lighting conditions **in the background**") and the light illuminating the dress ("On a scale from 1 to 5, where 1 is cool and 5 is warm, please rank the light **illuminating the dress**"). They were then asked to characterize the light in the background, and the light illuminating the dress, by checking off any of a number of possible verbal

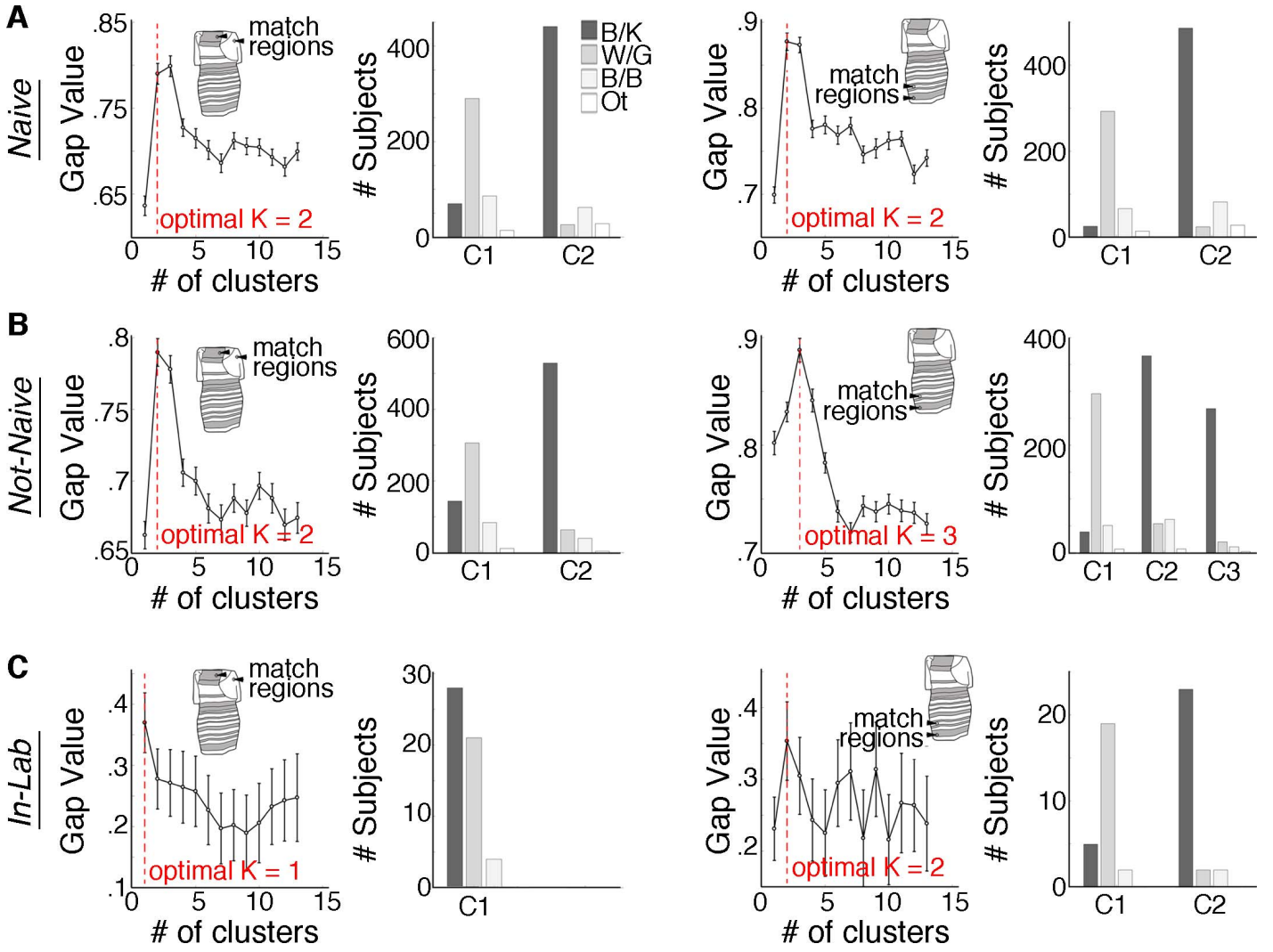

Figure 4. *K*-means clustering solutions for naïve, non-naïve, and in-lab subjects. The *k*-means clustering assessments were performed for the subsets of subjects who (A) had never seen the dress before participating in the study ($N = 1{,}017$; top row), (B) had seen the dress before ($N = 1{,}183$; middle row), and (C) who were tested under controlled conditions in the laboratory ($N = 53$; bottom row). Conventions are as in Figure 2.

descriptors from a list ("The **lighting in the background** is... Check all that apply/The light **illuminating the dress** is... Check all that apply": dim, dark, cool, blueish, bright, warm, yellowish, glaring, blown out, washed out, reddish, greenish, purplish, iridescent).

Color matching

Each image was presented a second time (again, for 15 s), and this time subjects were prompted to make color matches to four regions of the dress (Figure 1 inset, arrows i, ii, iii, iv), using a color-picker tool comprising a complete color gamut: "Please adjust the hue (color circle) and brightness (slider bar) to match the pixels you see in the image."

For Experiments 1 and 3, in the first half of the experiment each image was shown for 15 s and then subjects were prompted to perform the color-naming task and the lighting-judgment task. In Experiment 2, subjects were not asked to perform the lighting judgment task. The image was on the screen continuously while the subjects responded (this was not designed as a test of color memory). In the second half of the experiment, each image was shown a second time, and after 15 s, subjects were prompted to perform the color-matching task (the image remained on the screen continuously while the subjects performed the task). Between the presentation of the first and second images, we collected basic demographic information. Between the presentations of the last two images, we asked subjects about the environment in which they were completing the study. We also asked whether they had viewed the original dress image prior to this study,

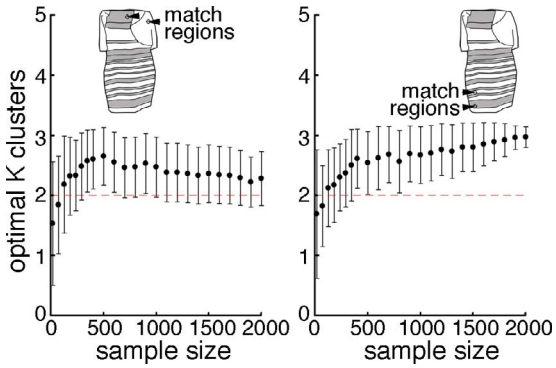

Figure 5. Power analysis. Plots show the results of *k*-means clustering (as in Figure 2) for a range of sample sizes (color-matching data randomly sampled from the online subject pool). For each sample size tested, the average gap value derived from 100 bootstraps is plotted; error bars show the standard deviation. Left panel corresponds to upper dress-region matches; right panel, to lower dress-region matches. The variance around the predicted *k* decreases and the predicted *k* increases with increasing numbers of samples. That the optimal $k = 2$ or 3, and not 1, becomes significant with about 125–180 subjects, and plateaus around 500 subjects.

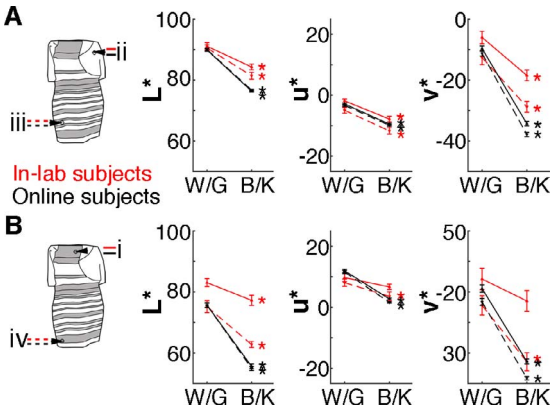

Figure 6. Color matches made by B/K reporters and W/G reporters differ in lightness and hue, under controlled conditions and online. (A) Matches for the blue/white regions. (B) Matches for the black/gold regions. Plots show the average lightness and hue components (CIELUV 1976: L*, u*, v*) of color matches made by subjects reporting B/K or W/G, tested online (black lines; $N = 1{,}174$; 770 B/K, 404 W/G) and under controlled viewing conditions (red lines; $N = 49$; 28 B/K, 21 W'/G). Solid lines show data from the upper regions of the dress, and dashed lines show data from the lower regions; error bars show 95% confidence intervals. Asterisks show cases for which B/K and W/G matches differed significantly (paired *t* tests): In the online data, matches differed in all three color dimensions (u*, or "red–greenness"; v*, or "blue–yellowness"; and L*, or Luminance; *p* values < 0.001); and in the in-lab data, matches differed in all but the u* and v* dimensions of one region (i; *p* values < 0.001 for all paired *t* tests, except u* and v* of region i: $p = 0.06$, 0.2).

and if so, whether they had experienced multiple percepts of the dress (i.e., switching).

The first image shown was always the original dress photograph (courtesy of Cecilia Bleasdale), but the order of the subsequent two images differed between the two conditions. In addition, all subjects were queried on their perception of a blurry version of the original image.

Experiment 1:

- Order A: Report colors and lighting for original dress image, cue-rich scene (cool), cue-rich scene (warm); make color matches for original dress image, cue-rich scene (cool), cue-rich scene (warm); report colors of the blurry dress.
- Order B: Report colors and lighting for original dress image, cue-rich scene (warm), cue-rich scene (cool); make color matches for original dress image, cue-rich scene (warm), cue-rich scene (cool); report colors of the blurry dress.

Experiment 2:

- Order: Report colors for original dress image; make color matches for original dress image.

Experiment 3:

- Order A: Report colors and lighting for original dress image, cue-rich scene (cool), cue-rich scene (warm), uniform surround (cool), uniform surround (warm), skin-tint only (cool), skin-tint only surround (warm); make color matches for original dress image, cue-rich scene (cool), cue-rich scene (warm), uniform surround (cool), uniform surround (warm), skin-tint only (cool), skin-tint only (warm); report colors of the blurry dress.
- Order B: Report colors and lighting for original dress image, cue-rich scene (warm), cue-rich scene (cool), uniform surround (warm), uniform surround (cool), skin-tint only (warm), skin-tint only surround (cool); make color matches for original dress image, cue-rich scene (warm), cue-rich scene (cool), uniform surround (warm), uniform surround (cool), skin-tint only (warm), skin-tint only surround (cool); report colors of the blurry dress.

Data analysis

Analyses are described in the legends and Results section. All statistical analyses were conducted using MATLAB. A value of $p < 0.05$ was considered statistically significant.

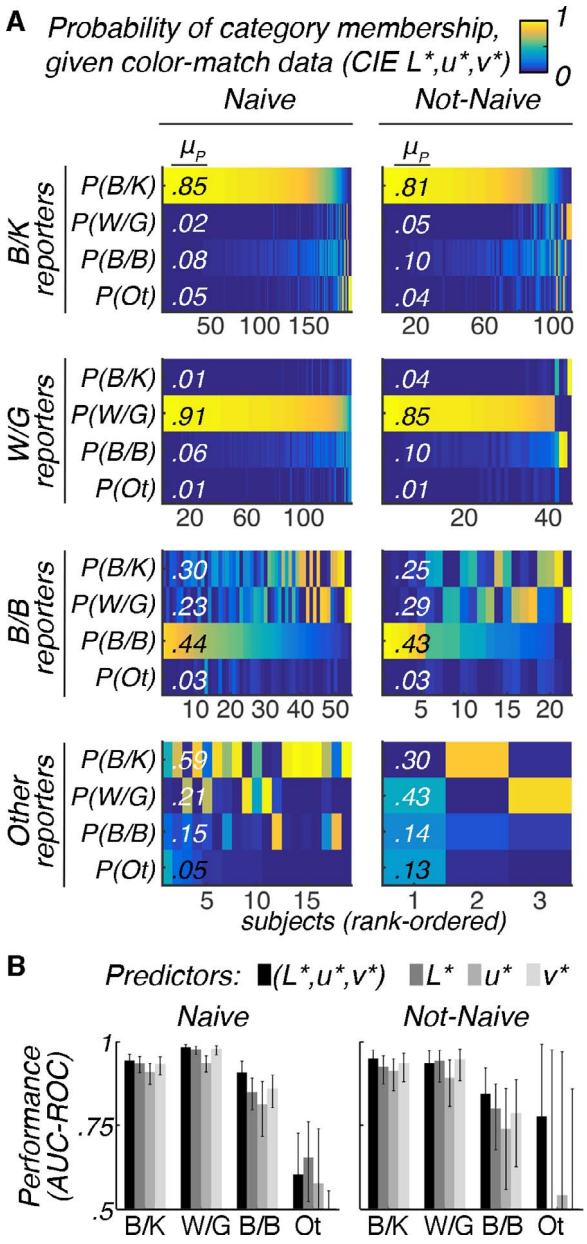

Figure 7. Categorical verbal reports can be predicted from color matches. Multinomial logistic regression was used to build nominal-response models (classifiers). Four models were generated: The full model was fitted using the L*, u*, and v* components of subjects' color matches (to all four match regions) as predictors; three additional models were fitted using either the L*, u*, or v* component of subjects' matches as the predictors. Models were fitted with responses from a subset of the online subjects (half the subjects from Experiment 2, $N_{train} = 549$) and tested on responses from the left-out subjects ($N_{test} = 547$). (A) Predicted probability of category membership for the full model. Each panel contains the results for data from individual (left-out) subjects, grouped by the verbal label they used to describe the dress (ground truth) and whether they had seen the dress prior to the study (Naïve vs. Not-Naïve). Each thin vertical column within a panel shows the results for a single subject: →

Results

#TheDress (Figure 1, left) is a rare image that elicits striking individual differences in color perception (Gegenfurtner et al., 2015; Lafer-Sousa et al., 2015). Although the pixels that make up the dress are (in isolation) light blue and brown, most observers queried through social media reported seeing the dress as either blue/black (B/K) or white/gold (W/G; Rogers, 2015). A minority of subjects (~10%) reported seeing the dress as blue/brown (B/B) (Lafer-Sousa et al., 2015).

Categorical perception of the dress: True or false?

Subjects were asked to identify the colors of four regions of the dress (i–iv; Figure 1, left). The three-dimensional (CIE L*,u*,v*) coordinates of the color-matching data were compressed to one dimension using principal-component analysis (Lafer-Sousa et al., 2015). Subjects' color matches for the brown regions of the dress (i, iv) are plotted against their matches for the blue regions of the dress (ii, iii), and are highly correlated (density plots and contours were created in MATLAB using *scatplot*[1]). Moreover, the correlation shows two peaks, suggestive of two underlying categories. This pattern of results was consistent for both subjects who had and had not seen the image previously (Figure 1B, 1C). The peaks in the population density plots corresponded well with the categorical color descriptions provided by the participants: Figure 1D shows the color matches made by subjects

←
The colors of each row in the column represent the predicted probability that the subject used each of the categorical labels (B/K, W/G, B/B, Other); each column sums to 1. Subjects are rank-ordered by the predicted probability for the ground-truth class. The average predicted probabilities for each response category are denoted μ_P. (B) Bar plots quantifying classification performance (the area under the receiver operating characteristic curves, computed using the true and false positive rates), by category, for each of the four models. Error bars indicate 95% confidence intervals. Values greater than 0.90 indicate excellent performance; values between 0.75 and 0.90 indicate fair to good performance; values between 0.5 and 0.75 indicate poor performance. We compared the accuracy of the various models against each other using MATLAB's testcholdout function: The L*-only model performed no better than the u*-only model (Naïve: $p = 0.5$; Not-Naïve: $p = 0.8$) or the v*-only model (Naïve: $p = 0.5$; Not-Naïve: $p = 0.3$). The full model was more accurate than the L*-only model, but only among Naïve subjects (Naïve: $p < 0.001$; Not-Naïve: $p = 0.09$). True positive rates (sensitivity) for all four models are provided in Table 1.

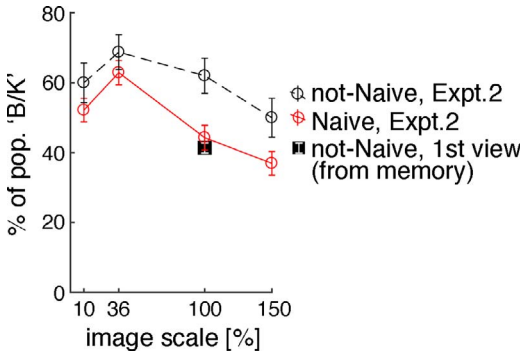

Figure 8. Image scale and prior exposure affect perception of the dress's colors. In Experiment 2 we varied the presentation scale of the image (different subjects saw different scales; 10%, 36%, 100%, 150%). Unlike in Experiment 1, subjects were never shown an image in which the lighting cues were disambiguated. Open circles show the percentage of subjects who reported B/K at each scale. The red line shows data from subjects who had not seen the dress prior to participating in the study (10%, $N = 207$; 36%, $N = 194$; 100%, $N = 192$; 150%, $N = 195$). The black line shows data from subjects who had seen the dress prior to the experiment (10%, $N = 80$; 36%, $N = 80$; 100%, $N = 100$; 150%, $N = 78$). Viewing the image at a reduced scale and having prior experience with the image both increased the proportion of subjects who reported B/K, and the two factors interacted (ANOVA performed with bootstrapped data from the open circles showed main effects of scale, $p < 0.001$, experience, $p < 0.001$, and their interaction, $p < 0.001$). Subjects who had seen the image before also reported on the colors they perceived when they first saw it, recalled from memory (solid black square; data from Not-Naïve subjects in Experiments 1 and 2, $N = 1{,}037$). The distribution of verbal color reports corresponding to the first time that Not-Naïve subjects viewed the image differed from the distribution of reports the same subjects provided in response to our presentation of the image, confirming that many had flipped (chi-square test of independence: $X^2 = 731$, $p < 0.001$). Error bars show standard errors (bootstrapping, sampling with replacement).

who reported a B/K percept (left panel) or a W/G percept (middle panel). Some subjects reported B/B. These subjects made color matches that were intermediate to the two main categories (Figure 1D, right panel). To quantitatively test the hypothesis that the dress is viewed categorically, we performed a k-means clustering assessment on the color-matching data.

There are several methods for estimating the optimal number of k clusters (groups) in a distribution, but most are constrained to assessing solutions of two or more clusters. To test our hypothesis, we need some way of assessing the relative goodness of clustering for a single-component ($k = 1$ cluster) versus a $k > 1$ component model. To do so, we clustered the data, varying the number of clusters ($k = 1, 2, \ldots, 13$), then used the gap method to assess the outcomes and identify the optimal K clusters (Tibshirani, Walther, & Hastie, 2001). The gap statistic is estimated by comparing the within-cluster dispersion of a k-component model to its expectation under a single-component null, and seeks to identify the smallest k satisfying $\text{Gap}(k) \geq \text{Gap}(k+1) - \text{SE}(\text{Gap}(k+1))$. For color matches made to both upper (Figure 2A–2C) and lower (Figure 2D–2F) regions of the dress, the single-component solution was rejected in favor of two or three clusters, confirming the suspected nonunimodality of the underlying population distribution.

The bar plots (Figure 2B, 2E) show the distribution of the different color terms assigned to each cluster, for the optimal k solution returned by the gap analysis. The clustering algorithm assigned the majority of W/G reporters' matches to Cluster 1 (in both the upper- and lower-region analyses) and the majority of B/K reporters' matches to Cluster 2 (upper-region analysis) or Clusters 2 and 3 (lower-region analysis). Matches made by subjects who described the dress as B/B or other colors outside of the main categories (other) were distributed more evenly across the clusters, even when three clusters were returned. Each thin band in the tapestries (Figure 2C, 2F) corresponds to the color matches made by a single subject, providing a visual snapshot of the success of the clustering algorithms in separating W/G and B/K reporters. These results show that perception of the dress in the population is categorical, not continuous, and reject the idea that reports of categorical perception are an artifact caused by a forced choice.

Categorical perception of the dress: How many categories?

The results of Figure 2 suggest that the underlying population may comprise two or three distinct categories. In our initial report, we argued that subjects who report the dress as B/B constituted a distinct third category, intermediate between the two main categories. Figure 3 (top row) shows the spatial relationship of the optimal k cluster centroids (x) and the color-matching distributions (contours) for subjects grouped by their verbal report, for an analysis of the upper match regions. The cluster centroids coincide with the center of the color-matching data for W/G and B/K subjects (xs fall inside the contours; contours and centroids obtained with different halves of the data). Color matches made by subjects who reported B/B fell between these centroids (Figure 3, top right panel). These results suggest that the B/B report does not reflect a distinct category. Figure 3 (bottom row) shows the relationship between the optimal k cluster centroids for the bottom match regions, which returned three clusters. But none of the centroids fell within the contour capturing color matches made by B/B subjects (Figure 3, bottom right panel). These results show that the third category, when

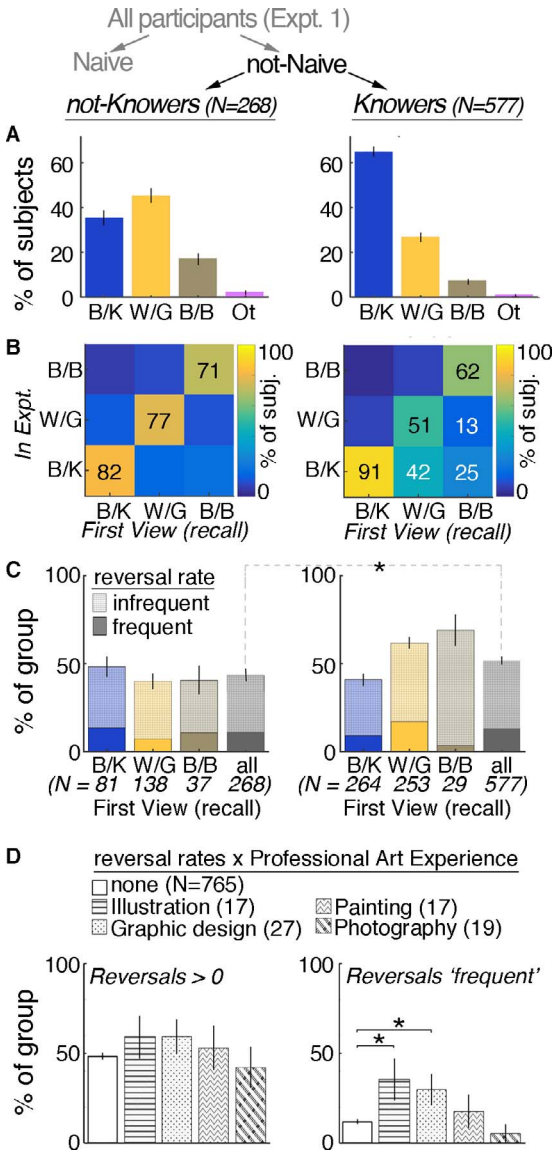

Figure 9. Familiarity with the dress image affects subsequent viewings. (A) The distribution of color reports (Experiment 1; image scale always 36%) from subjects who had seen the image before ($N = 845$) and had knowledge of the dress's colors in real life ("Knowers," $N = 577$) differed from the distribution of reports from subjects who had seen the image and did not know the dress's real colors ("Not-knowers," $N = 268$; chi-square test for difference of proportions: $p < 0.001$, $X^2 = 66$). (B) Color reports of the dress during the experiment as a function of how subjects first perceived the dress (recalled from memory). Results for Not-knowers are predominantly along the $x = y$ diagonal, reflecting dominance of initial stable state (though the presence of deviations was still significant: chi-square test of independence: $X^2 = 277$, $p < 0.001$). Results for Knowers showed more substantial deviations from the diagonal, particularly for subjects who first saw W/G or B/B, reflecting a weakening of initial state and an increased dominance of B/K state (the color of the dress in real life; chi-square test of independence: $X^2 = 394$, $p < 0.001$). (C) Quantification of self-reported reversal rates for Knowers and Not-knowers; percentages computed within class (B/K, W/G, B/B, Other; class corresponded to first-view percepts from memory). Subjects were asked: "In your viewings prior to this study, did your perception of the dress colors ever change? Y/N" and "How often did you see it change? Frequently/Infrequently/Never." The proportion of subjects who reported changes was higher among Knowers than Not-knowers (gray asterisk; chi-square test for difference of proportions: $p = 0.036$, $X^2 = 4.4$). Error bars indicate standard errors (bootstrapping, sampling with replacement). (D) Quantification of self-reported reversals as a function of professional art experience. Left panel shows the proportion of subjects who reported at least one reversal, sorted by professional art experience. Right panel shows the proportion of those subjects who reported that reversals were frequent. Compared to non-artists, reports of frequent reversals were significantly higher among subjects who indicated having professional art experience in the fields of illustration ($p = 0.004$, two-proportion z test) and graphic design ($p = 0.005$). Error bars indicate standard errors (bootstrapping, sampling with replacement).

evident, is a subgroup of the population of observers who describe the dress as B/K. In addition to the gap analysis, we applied the silhouette clustering criterion and the Calinski–Harabasz clustering criterion; these methods do not allow for a single-component solution. They returned an optimal solution of two clusters, for both the upper and lower match regions.

Does prior exposure to the image change the number of categories manifest in the population? The color-matching data for subjects with versus without prior exposure were similar (Figure 1), suggesting that prior exposure had no impact on the number of categories in the population. Figure 4 shows gap-statistical tests of the color-matching data to establish this conclusion. The optimal number of clusters for subjects with no prior exposure ($N = 1{,}017$), for either the top or bottom regions of the dress, produced an optimal cluster number of two (Figure 4A); the only color-matching data that produced more than two optimal clusters were those obtained on subjects with prior exposure tasked with matching the lower part of the dress ($N = 1{,}183$, three clusters, Figure 4B). The results of the analyses carried out using the silhouette clustering criterion and the Calinski–Harabasz clustering criterion returned two clusters, regardless of whether the matches came from subjects with prior experience.

Together with the qualitative evaluation of the color-matching data we reported previously (Lafer-Sousa et al., 2015), these results show that the categories reported in social media reflect true categories in the population and are not a result of the way the question was posed.

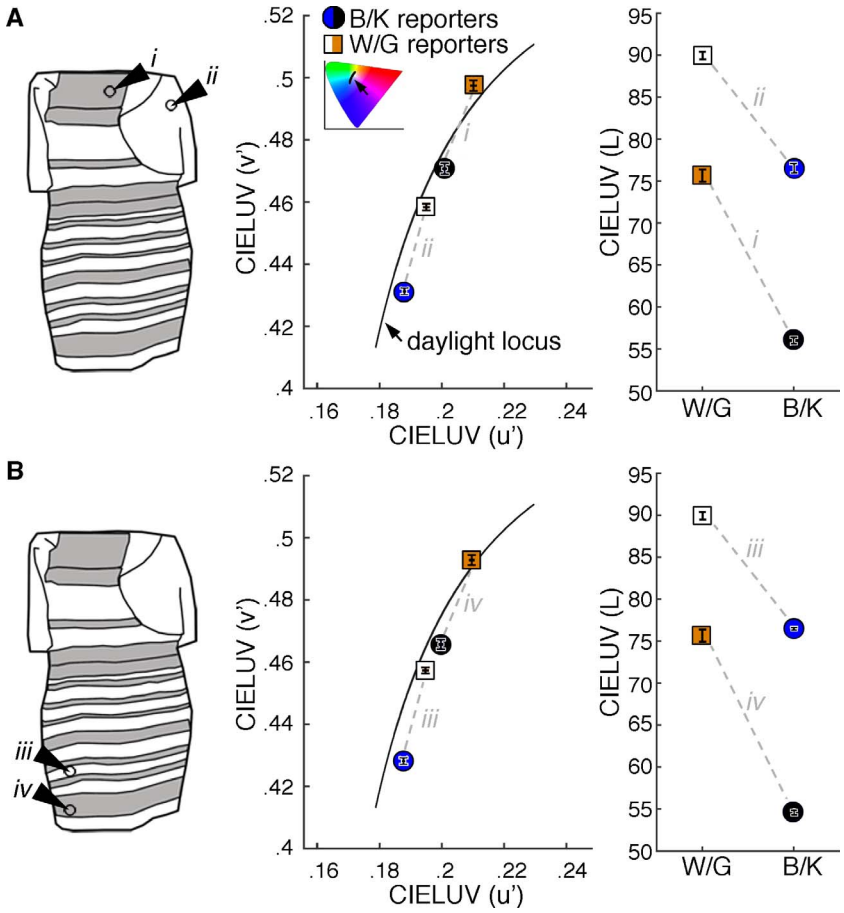

Figure 10. Color matches made by B/K and W/G reporters track the daylight locus. Chromaticity matches for (A) the top regions of the dress and (B) the bottom regions of the dress. Left graphs: Mean hue (CIE u′, v′) of color-matches made by online subjects ($N = 1{,}174$; B/K = 770; W/G = 404). Right plots: Mean lightness of matches (CIE L). Error bars show the 95% confidence interval of the mean. Data were grouped by the verbal report made by the subjects: squares for W/G and circles for B/K. The color of the symbol corresponds to the color term used by the subject (blue, black, white, or gold). Black line shows the daylight locus. Inset = CIE 1976 u′v′ color space with the daylight locus (arrow).

Categorical perception of the dress: Power analysis

The categorical nature of the population distribution appears, on first inspection, to contradict reports that the true population distribution is continuous. The discrepancy may be resolved by considering the large differences in the number of subjects in the different studies: Conclusions of a continuous distribution were made using data from less than two dozen subjects, whereas the present analysis depends on data from several thousand subjects. How many subjects are necessary to uncover the true population distribution? To address this question, we performed a power analysis by computing the optimal k using subsamples of the data we collected, and then bootstrapping (Figure 5). The variance around the predicted k decreases and the predicted k itself increases with increasing numbers of samples. That the optimal k is 2, and not 1, becomes significant with about 125–180 subjects.

The gap statistic obtained for all subjects ($N = 2{,}200$), for the upper match regions, yielded comparable values for two and three clusters (Figure 2); in this case, the optimal gap value is considered to be 2, because the gap method favors the smallest number of clusters satisfying the method's criterion. The gap statistic for the lower match regions was clearly distinguished as three clusters. This difference between the upper and lower match regions accounts for the difference in the results of the power analysis: For the upper regions, the optimal k converges between two and three clusters (and the error bars remain large even at large sample sizes); for the lower regions, it converges on three clusters and the error bars get very small at large sample sizes. The additional cluster identified using color matches for the bottom region of the dress correspond not to a discrete B/B category but rather to a subdivision of the B/K category (Figure 3B; two centroids fall within the B/K reporters).

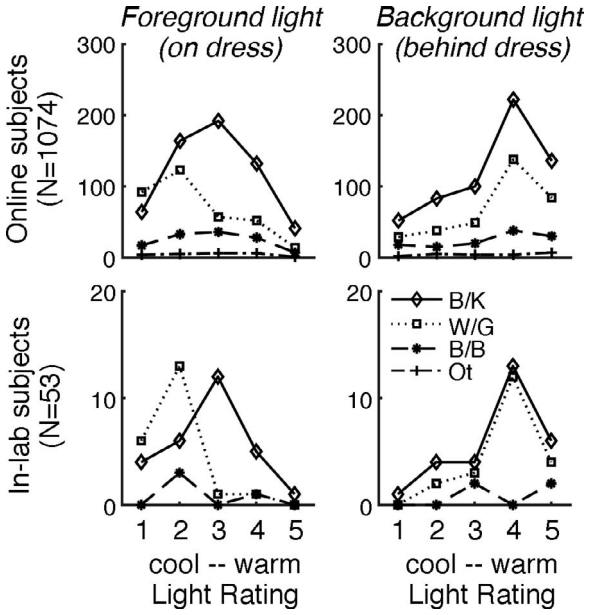

Figure 11. Temperature ratings of the light shining on the dress systematically differ as a function of percept. Subjects (top row: 1,074 online, bottom row: 53 in-lab) were asked to rate their impression of the light shining on the dress and the light illuminating the background, on a scale of 1–5 (cool to warm). Plots show illumination ratings for the light on the dress (left panels) and the light in the background (right panels), grouped by verbal report of the dress's colors (B/K, W/G, B/B, Other). Subjects' ratings of the light on the dress systematically differed as a function of percept (two-sample t test comparing the ratings provided by subjects who saw the dress as B/K vs. those who saw W/G; online: $p < 0.001$; in-lab: $p = 0.002$). Subjects' ratings of the background light did not differ as a function of percept (online: $p = 0.2$; in-lab: $p = 0.5$).

Our estimate of the number of subjects required to adequately assess the underlying population distribution was made using data obtained online; there is much higher variability in the viewing conditions and subject pool for experiments conducted online versus in a lab. It is possible that the number of subjects needed to determine the true underlying population distribution would be lower for data obtained in a lab, where the viewing conditions can be better controlled. In our prior study, we collected data from 53 subjects under controlled lab conditions. A k-means clustering analysis of these data neither rejected nor confirmed a single-component model (Figure 4C). These results suggest that conducting the experiments in a lab confers no benefit in uncovering the population distribution.

Categorical perception of the dress: Comparing results obtained online and in a lab

The results in Figure 4C show that even under controlled viewing conditions, samples of more than 53 participants are needed to reliably uncover the true population distribution. Nonetheless, the data collected in the lab showed trends consistent with two underlying categories of observers. First, qualitative assessment of the color-matching plots (Lafer-Sousa et al., 2015) shows evidence of two clumps. Second, the optimal k for one of the two sets of regions (iii, iv) was 2, even if this optimal value is not strongly distinguished from other values of k (Figure 4C, left panel). That the optimal k is 2 becomes clearer when the analysis is run on data combining color matches for all regions tested, essentially doubling the data set, which returns an optimal k of 2 (data not shown). Third, the relative distribution of W/G to B/K observers among participants without prior exposure was about the same for subjects tested online versus in a lab (Lafer-Sousa et al., 2015). And fourth, the average chromaticity of the color matches made by subjects in the lab was consistent with those made by subjects online (Figure 6). Regardless of whether the data were obtained online or in a lab, the results showed the same pattern: Compared to B/K subjects, W/G subjects reported not only higher luminance but also higher values of u* (redness) and v* (yellowness), for all four regions tested. The strongest changes were in the luminance and v* dimensions. The comparability of data collected online versus in a lab is consistent with the idea that the factors that determine how one sees the dress are relatively high level, divorced from the specific low-level conditions of viewing (such as the white balance, mean luminance, and size of the display).

Categorical perception of the dress: Predicting color terms from color matches

The results in Figures 1–6 support the conclusion that the color of the dress photograph is resolved as one of two dominant categories, consistent with the initial social-media reports. As an additional test of the hypothesis, we performed a multinomial logistic regression analysis to test the extent to which we could predict the colors a person would use to describe the dress, given the color matches the person makes. If, for example, B/B reporters represent a distinct or stable perceptual category, then a classifier (trained on independent data) should be able to distinguish B/B reporters from B/K and W/G on the basis of color matches alone. We generated four models: a full model, which used the L*, u*, and v* components of subjects' matches as the predictors, and three additional models which used either the L*, u*, or v* component of subjects' matches as the predictor. The models were trained and tested with independent data ($N_{train} = 549$, $N_{test} = 547$). Figure 7A shows the prediction outcomes for the full model. Panels show the results grouped by

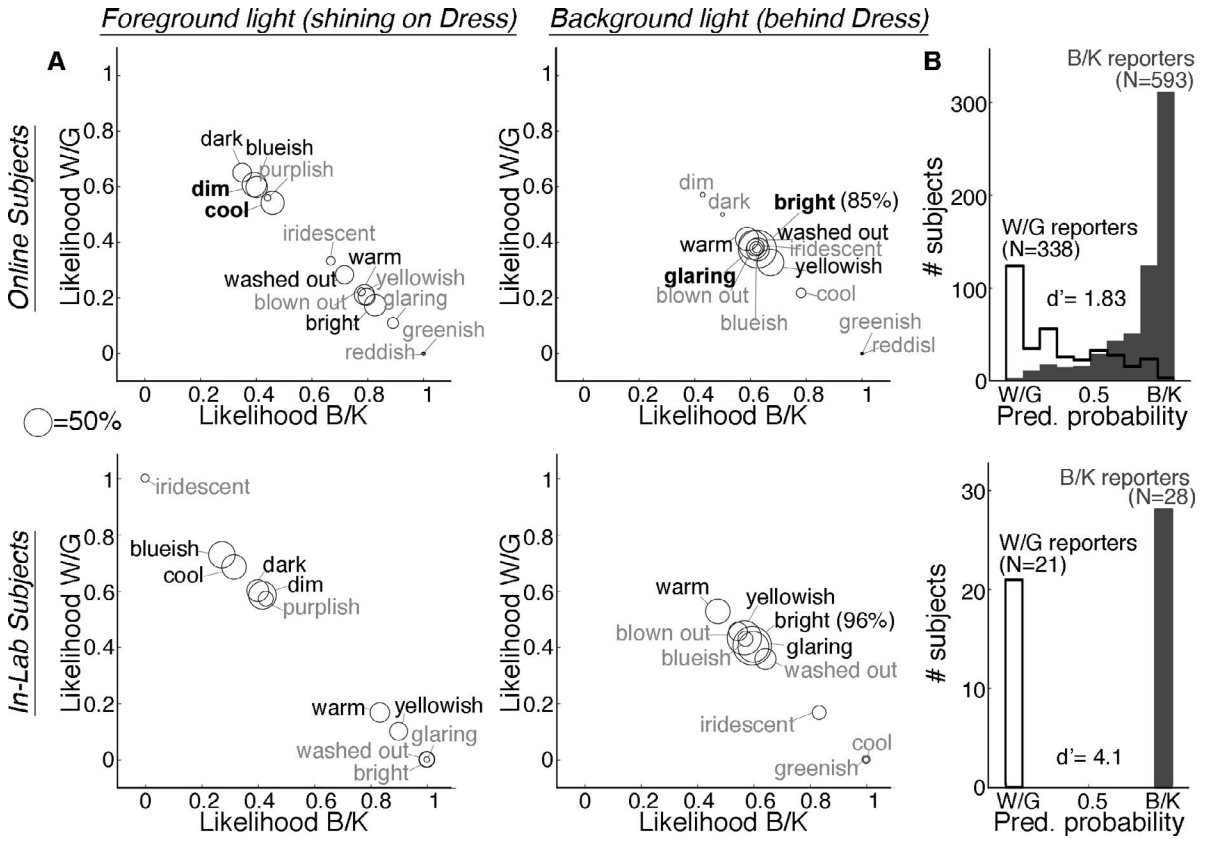

Figure 12. Subjects' color percepts of the dress are predicted by their inference of the lighting conditions. Subjects (top row: 1,074 online; bottom row: 49 in-lab; only subjects who reported either W/G or B/K are included in this analysis) were asked to characterize the light shining on the dress and the light illuminating the background, by checking off any of a number of possible verbal descriptors from a list (*dim*, *dark*, *cool*, *blueish*, *bright*, *warm*, *yellowish*, *glaring*, *blown out*, *washed out*, *reddish*, *greenish*, *purplish*, *iridescent*). (A) For each word in the list, the likelihood of being B/K = (# of B/K reporters who used the term)/(# of B/K + # of W/G who used the term). The diameter of the bubble reflects the proportion of people in the population who used the term: (# of B/K who used it + # of W/G who used it)/(# of B/K reporters + # of W/G reporters). Inset key = 50%. Bubble plots for the light shining on the dress (foreground; left panels) and the light in the background (right panels; top row: online subjects; bottom row: in-lab subjects). (B) Classification histograms for a binary logistic regression where the lighting descriptors were used as predictors to distinguish B/K from W/G reporters—online subjects: correct rate = 84% (85% for B/K, 80% for W/G), $d' = 1.83$, $R^2 = 0.52$; in-lab subjects: correct rate 100%, $d' = +\infty$, $R^2 = 1$. A test of the full model against a constant-only model was statistically significant, indicating that the predictors (the verbal descriptors of the lighting conditions) as a set reliably distinguish between B/K reporters and W/G reporters (online subjects: $X^2 = 583$, $p < 0.001$; in-lab subjects: $X^2 = 66.9$, $p = 0.037$).

prior experience with the image (Naïve vs. Not-Naïve) and ground-truth verbal label (B/K, W/G, B/B, Other). Among participants who reported B/K, the average predicted probability of B/K category membership was 0.85 for subjects with no prior experience (Figure 7A, top left) and 0.81 for subjects with prior experience (Figure 7A, top right). These probabilities were such that 92% of B/K reporters were classified B/K (93% of those without prior experience, 89% of those with prior experience); and 98% of W/G reporters were classified W/G (100% of those without prior experience, 91% of those with prior experience; Table 1). The model was less successful in its classification of B/B reporters, but above chance: 51% of B/B reporters were classified B/B (54% of those without prior experience, 45% of those with prior experience). The classifier failed to classify all Other reporters as Other. Overall, the true positive rate was 85%. We further quantified model performance by computing the area under the receiver operating characteristic (ROC) curve for each categorical label. This quantity reflects both the true and false positive rates (Figure 7B). Values greater than 0.90 indicate excellent performance (which we observed for B/K and W/G reporters), while values between 0.75 and 0.90 represent fair to good (observed for B/B reporters) and values between 0.5 and 0.75 represent poor performance (observed for Other reporters). Values below 0.5 represent failure.

Do the model results contradict the conclusions drawn in Figures 2 and 3 by providing evidence for

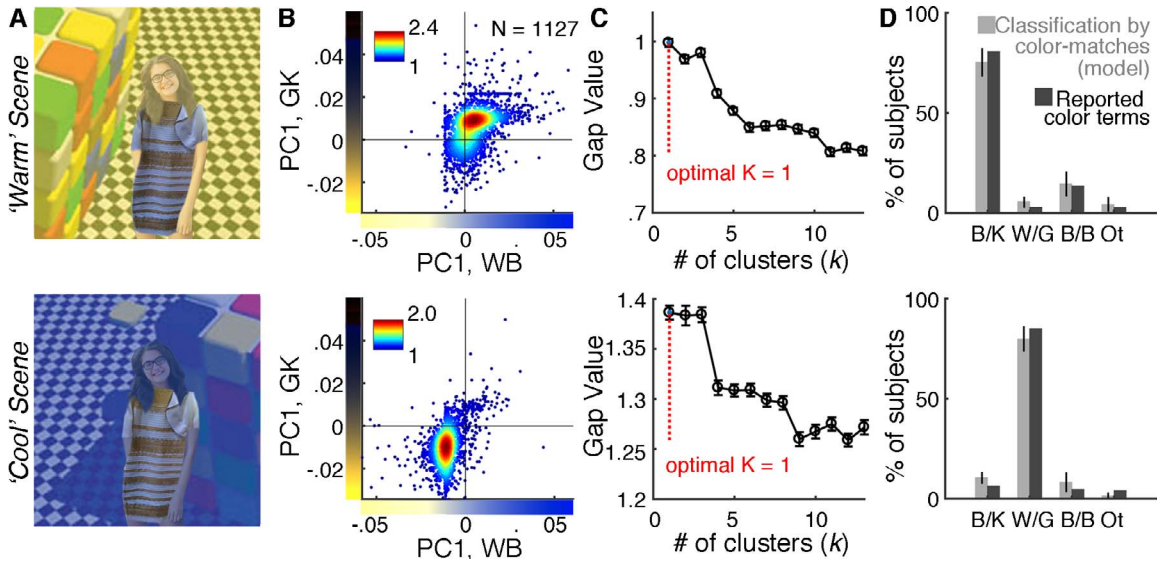

Figure 13. When the dress is embedded in scenes with overt cues to the illumination, it is perceived in a way that can be predicted from the overt cues: W/G when a cool illuminant is cued, B/K when a warm illuminant is cued. (A) The dress was digitally embedded in simulated contexts designed to convey warm illumination (top row) or cool illumination (bottom). Cues to the illuminant are provided by a global color tint applied to the whole scene, including the skin of the model, but not the pixels of the dress. (B) Distribution of subjects' ($N = 1{,}127$; 1,074 online, 53 in-lab) color matches. Conventions as in Figure 1. (C) Results of *k*-means clustering assessments of the matching data from (B). Conventions as in Figure 2. The analysis favored a single-component model (optimal $k = 1$ cluster) for both the warm and cool scene distributions. (D) Distribution of categorical percepts observed (dark-gray bars) and the distribution predicted from the multinomial classifier (light-gray bars; see Figure 7). Error bars show 95% confidence intervals. Photograph of the dress used with permission; copyright Cecilia Bleasdale.

three categories? While the true positive rates for the B/B observers (51%) are above chance (chance = 25%) and the ROC analysis shows fair to good performance, only ~25% of the correctly classified B/B reporters were classified accurately with high predicted probability. Furthermore, among those misclassified, most were classified with strong confidence as being either B/K or W/G (Figure 7A, third row). There was a delay of a few minutes between when subjects provided verbal reports about the dress color and when they did the color-matching experiment; the perception of the dress could have switched during this time, as spontaneously reported by some subjects. That many B/B people were confidently classified as either B/K or W/G is consistent with the hypothesis that these subjects may have switched percepts between when they gave their verbal report and when they gave their color match. Together with the results in Figures 2 and 3, the classifier results suggest that the B/B designation is not a distinct category. Instead, we interpret the results to indicate the B/B category as a transient state between B/K and W/G, which would be consistent with the properties of a bistable phenomenon.

		True positive rate				
Model predictor	Subjects	All	B/K	W/G	B/B	Other
(L*, u*, v*)	Naïve	86%	93%	100%	54%	0%
	Not-Naïve	83%	89%	91%	45%	0%
L*	Naïve	79%	96%	96%	2%	0%
	Not-Naïve	78%	91%	89%	5%	0%
u*	Naïve	78%	92%	89%	24%	0%
	Not-Naïve	79%	89%	89%	18%	0%
v*	Naïve	80%	89%	99%	30%	0%
	Not-Naïve	81%	86%	96%	36%	0%

Table 1. True positive rate of multinomial classifiers trained to predict verbal color reports from color matches (see Figure 7). Four models were generated: The full model was fitted using the L*, u*, and v* components of subjects' color matches as predictors; three additional models were fitted using either the L*, u*, or v* component of subjects' matches as predictors. Models were fitted with responses from a subset of the online subjects (half the subjects from Experiment 2, $N_{train} = 549$) and tested on responses from the left-out subjects ($N_{test} = 547$). The table shows the true positive rates—(# of true positives)/(# of true positives + # of false negatives)—for each model, broken down by subjects' prior experience with the image (Naïve, Not-Naïve) and their verbal label (B/K, W/G, B/B, and Other).

When the model was fitted using only the L*, u*, or v* components of subjects' color matches as the predictors, the classification performance remained high (Table 1; Figure 7B) and did not differ as a function of which component was used as the predictor: The L*-only model performed no better than the u*-only model (no prior experience: $p = 0.5$; prior experience: $p = 0.8$) or the v*-only model (no prior experience: $p = 0.5$; prior experience: $p = 0.3$). The full

model was more accurate than the L*-only model, but only among subjects with no prior experience ($p < 0.001$; prior experience: $p = 0.09$). This provides additional support for the contention that B/K and W/G reporters are differentiated by both the lightness and the hue of the matches they select.

Categorical perception of the dress: Switching perception from one category to another

The evidence presented in Figures 1–7 strongly suggests that the dress image is an ambiguous image that the visual system can interpret as one of two mutually exclusive categorical percepts. When viewing ambiguous shape images such as the Necker cube, subjects often report a change in their perception of the image from one stable state to the other. But it has also been shown that knowledge of the fact that the image can be perceived in different ways can have a profound impact on whether subjects see the image flip. The online data we obtained came from a diverse subject pool. Unlike most other online surveys of the phenomenon, participants were not recruited with links attached to media reports describing the dress. As a result, the image was entirely new to many of the participants in our study. Moreover, of those who had previously encountered the image, many had no knowledge of the actual color of the real dress, enabling us to test the extent to which subjects can change their perception of the dress and, if so, the impact on flip rates conferred by knowledge of the image's multistability.

We asked subjects who had previous experience with the image about their first encounter with it. The distribution of verbal color reports corresponding to the first time that these subjects viewed the dress (recalled from memory; Figure 8, filled symbol) showed a greater proportion of W/G than was reported by the same subjects in response to our presentation of the image ($X^2 = 731$, $p < 0.001$). Some of these subjects must have switched their perception of the dress's colors since their first viewing; the results show that experience with the image biased the population toward the B/K percept. The relative proportion of B/K first-encounter reports is comparable to findings from other surveys and matches the proportion recovered in experiments of subjects without experience of the image who were shown it at 100% size (the solid symbol overlaps the open red circle; Figure 8). As we showed previously, reducing the size of the image from its native size on the Internet also biased the population toward the B/K percept. These two factors (image size and experience) interact (ANOVA performed with the data from the Figure 8; open circles confirmed main effects of scale, $p < 0.001$, experience, $p < 0.001$, and their interaction, $p < 0.001$).

Among the 845 subjects with prior experience, 577 knew that the true dress color was B/K, allowing us to ask how knowledge alters perception. Knowledge of the dress's color in real life dramatically altered the ratio of B/K to W/G responses in the population in a direction that favored B/K (Figure 9A, 9B). Moreover, knowledge of the true colors increased reports of flipping between B/K and W/G (Figure 9C). Among people who first saw W/G yet reported knowing the true colors of the dress, 42% had switched to B/K (we confirmed that these individuals were not merely reporting the dress's true colors while continuing to perceive W/G by running their color matches through our classifier: Only 9% of them were classified as W/G; 76% were classified as B/K, 13% as B/B). Among people who first saw W/G and reported not knowing the true colors of the dress, only 14% had switched to B/K. These results show that knowledge of the dress's true colors affected whether people could see it flip.

The distribution of observers across categories was different between participants with and without knowledge of the true dress color (compare figure 1 of Lafer-Sousa et al., 2015, with Figure 9A, right panel; $X^2 = 21$, $p < 0.001$; all data obtained with the 36% image). Specifically, the proportion of B/K among participants with knowledge of the true colors (65%) was higher than the proportion of B/K among participants with no previous experience (54%). Curiously, among the population who had seen the dress previously but did not know its colors, we found roughly the same proportion of B/K as W/G observers (Figure 9B, left panel); this distribution was different than observed for participants with no prior experience, who were much more likely to see B/K ($X^2 = 23$, $p < 0.001$). Individuals with prior exposure but no knowledge of the dress's colors likely first saw the image in social media, where it was shown at a larger scale than we used in this set of experiments. We attribute the relatively higher levels of W/G reports among this group—even though the image we showed was at the smaller size—to the fact that their first view likely established a prior about the colors of the dress that had not been updated with any subsequent knowledge of its true colors.

On average, half of subjects reported experiencing the dress reverse at least once, while only 12% reported frequent reversals (Figure 9C). Given that reversal rates in multistable perception can be influenced by cognitive factors like personality, creativity, and attention, we examined the proportion of subjects reporting reversals as a function of their professional art experience (subjects could indicate professional experience with graphic design, illustration, photography, painting, and art history; Figure 9D). Although the proportion of

people who reported having seen the dress switch at least once did not differ as a function of art experience (Figure 9D, left), the proportion who reported frequent reversals was different across different art experience (Figure 9D, right). Compared to nonartists, reports of frequent reversals were 3 times higher among subjects who indicated having professional illustration experience ($p = 0.004$; two-proportion z test) and 2.5 times higher among those with professional graphic-design experience ($p = 0.005$).

Together, the analysis of the population responses of different categories of observers shows that (a) how you first saw the dress establishes a prior; (b) knowledge of the colors of the dress in real life updates this prior, biasing it toward B/K; (c) varying the image size systematically biases the percept (increasing image size increases W/G reporting; reducing image size increases B/K reporting); (d) experience with the image over time, independent of knowledge of the dress's true colors, biases the population toward B/K; and (e) reversal frequencies vary with professional art experience. These results uncover the important role played by both low-level perceptual features (such as image size) and high-level features (such as knowledge) in shaping how people perceive the colors of #TheDress, and add to a growing body of evidence that exposure to social media can change the colors we see.

What accounts for the different ways in which the dress colors are seen?

We have argued that the multistability of the image derives specifically from the fact that colors of the image align with the daylight locus (Brainard & Hurlbert, 2015; Conway, 2015; Lafer-Sousa et al., 2015). We hypothesized that in this context, multiple percepts become possible because the illumination cues in the image are ambiguous: Subjects may infer either a warm or a cool illuminant, and discount it accordingly. Consistent with this notion, color matches made by B/K and W/G reporters systematically shifted along the daylight locus, with B/K matches shifting away from the warm end of the locus (consistent with discounting a warm illuminant) and W/G matches shifting away from the cool end (consistent with discounting a cool illuminant; Figure 10A, 10B, left plots). W/G matches were also lighter on average than B/K matches, consistent with the idea that subjects are discounting not only chromatic biases in the illuminant but also lightness biases expected if they thought the dress was in shadow (Figure 10A, 10B, right plots). In our prior report, we tested the idea that illumination assumptions underlie the individual differences in color perception of the dress by determining how perception changes when the dress is embedded in a scene with disambiguated lighting (Lafer-Sousa et al., 2015). The results support the idea that observers who see B/K assume the dress is illuminated by a warm illuminant, while observers who see W/G assume that it is illuminated by a cool light source. Further evidence for this idea has been provided by others (Toscani, Gegenfurtner, & Doerschner, 2017; Witzel, Racey, & O'Regan, 2017).

To further test the hypothesis, we applied a classification algorithm to data collected in our initial survey (1,074 subjects online, 53 subjects in the lab), in which we asked people to explicitly report on the lighting conditions in the image (for the full questionnaire, see the supplementary material of Lafer-Sousa et al., 2015). We ran two experiments to assess subjective experience of the lighting conditions. First, we asked subjects to rate the illumination temperature on a scale of 1 to 5 for cool versus warm; second, we asked them to check off any of a number of possible verbal descriptors, including "dim," "dark," "bright," "warm," "cool," "blueish," and "yellowish" (see Methods and materials). Most subjects, regardless of their perception of the dress, reported the background illumination in the image to be warm (Figure 11, right panels; B/K and W/G reporter ratings did not differ, according to a t test—online subjects: $p = 0.2$; in-lab subjects: $p = 0.5$). But subjects who saw W/G differed from those who saw B/K in terms of their inference about the light on the dress itself: W/G percepts were associated with cool illumination, as if the dress were backlit and cast in shadow, while B/K percepts were associated with a warmer illumination, as if the dress were lit by the same global light as the rest of the room (Figure 11, left panels; t tests—online subjects: $p < 0.001$; in-lab subjects: $p < 0.001$). These analyses quantify results in our initial report and are consistent with other findings (Chetverikov & Ivanchei, 2016; Toscani et al., 2017; Wallisch, 2017; Witzel et al., 2017).

The results obtained using data on the warm/cool ratings were confirmed by an analysis of the descriptors that subjects used to characterize the lighting. W/G and B/K subjects were indistinguishable in the words they used to report on the illumination of the background (in Figure 12A the descriptors form one cluster in the right-hand bubble plots, where the most common descriptor was "bright") but showed strikingly different word choices when reporting on the illumination over the dress itself (descriptors form two clusters: Figure 12A, left-hand bubble plots), with words like "dim" and "cool" corresponding to higher likelihoods of W/G reporting and words like "warm" and "bright" to higher likelihoods of B/K reporting. The binary logistic regression using the lighting descriptors as predictors reliably classified B/K and W/G reporters (Figure 12B, classification histograms; online subjects: correct rate = 84%, $d' = 1.83$, $R^2 = 0.52$; in-lab subjects: correct rate = 100%, $d' = +\infty$, $R^2 = 1$) and outperformed a constant-

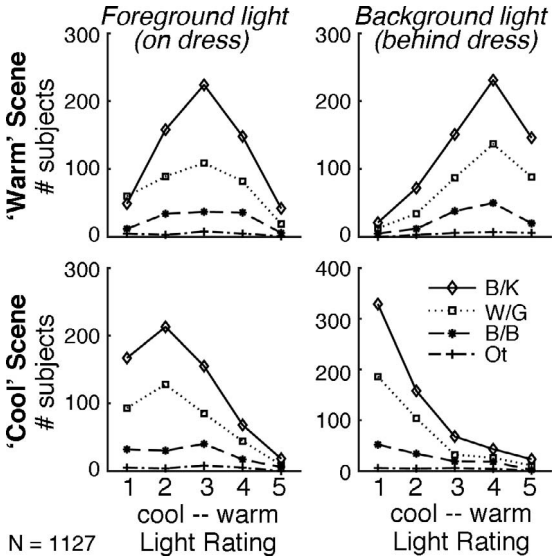

Figure 14. When the dress is embedded in scenes with overt cues to the illumination, the lighting is perceived in a predictable way. Subjects' ratings of the illumination in the simulated scenes from Figure 13 ($N = 1{,}127$; online = 1,074; in-lab = 53): warm illumination (top row) and cool illumination (bottom row). Conventions are as in Figure 11; data are grouped by the color terms that subjects used to describe the original dress image. Subjects' ratings of the foreground light did not differ as a function of initial percept ($p = 0.9$), nor did their ratings of the background light ($p = 0.8$; two-sample t tests comparing the ratings provided by subjects who originally reported B/K vs. those who originally reported W/G). Rating variance was higher for the original image (Figure 11) than for either test (cool: $p < 0.001$; warm: $p < 0.001$; F test), but similar for the tests ($p = 0.08$).

only model (online subjects: $X^2 = 583$, $p < 0.001$; in-lab subjects: $X^2 = 66.9$, $p = 0.04$). Note that, once again, the results obtained online were consistent with those obtained in the lab.

Quantification of color-matching data obtained from 1,127 people mostly online using cue-rich disambiguating stimuli show that for both of the disambiguating scenes (cool and warm illumination simulations; Figure 13A), the optimal k-means clustering solution is a single-component model (one cluster), consistent with the unimodal distribution of color terms reported under these conditions (Figure 13B–13D). When asked to rate the lighting conditions in the simulated scenes, subjects also conformed to ratings consistent with the lighting conditions cued (Figure 14). The results confirm that embedding the dress in scenes with unambiguous illumination cues resolves the individual differences in perception of the dress. The results also prove that the individual differences measured online, for the original dress image, are not simply the result of variability in viewing conditions.

What informs people's priors? Immediate prior exposure

Our results show that experience acquired over the medium term (knowledge of the true colors of the dress) affects perception. What about experience in the very short term? Presumably priors on lighting are updating constantly, weighted by the reliability of the data. We sought to test whether exposure to a disambiguated version of the image that was digitally manipulated to provide clear information about the illuminant affects how subjects see the original image.

We were able to address this hypothesis because we carried out two versions of Experiment 1 (with different participants; Figure 15). In one version, subjects provided color matches for the original dress image after being exposed to the image simulating a warm illuminant; in the second version, subjects provided color matches for the original dress image after being exposed to the image simulating a cool illuminant. We also conducted a separate experiment (Experiment 2) in which subjects were never exposed to the disambiguating stimuli. At the beginning of all experiments, subjects provided color terms for the original image. In our analysis we leveraged the discovery described previously (Figure 7), that color matches reliably predict verbal reports. We compared the verbal reports made by subjects at the beginning of the experiment with the verbal reports we predicted they would make, given their color matches, at the end of the experiment. If exposure to a disambiguated stimulus updates a prior about the lighting condition, the predicted verbal reports made on the basis of color-matching data should differ from the verbal reports made by the subjects for Experiment 1 but not Experiment 2; specifically, the predicted reports in Experiment 2 should be biased toward B/K when subjects were exposed to the warm scene and to W/G when exposed to the cool scene.

Figure 15A shows the results for the control case—no exposure to a disambiguating stimulus—and replicates the findings in Figure 1A: The density plot shows two strong peaks, corresponding to B/K and W/G reporters (Figure 15A is a subset of the data shown in Figure 1A). We deployed our classifier trained on independent data (Figure 7, full model) to categorize observers on the basis of the color-matching data they provided. The distribution of verbal reports predicted by the classifier (light-gray bars) is almost identical to the distribution of verbal reports that subjects provided (dark-gray bars; Figure 15D). Figure 15B shows the results of Experiment 1, Order A, in which a separate set of subjects ($N = 553$) viewed the simulated warmly lit scene immediately before they gave color matches for the original image. The distribution of verbal reports (obtained prior to color matching) is essentially indistinguishable from the distribution obtained in the control experiment (compare dark bars in Figure 15E

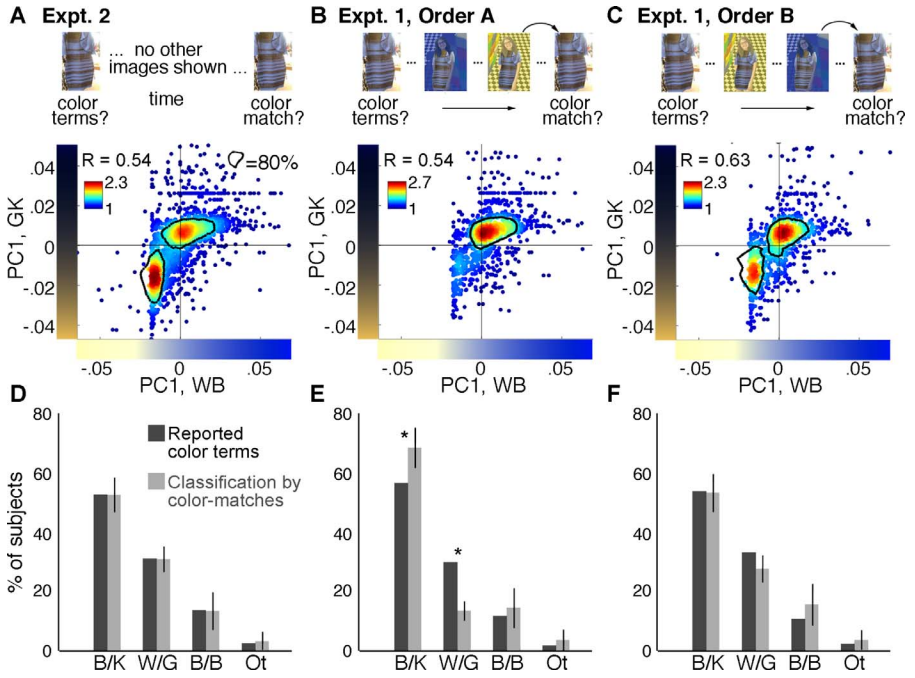

Figure 15. Priming with disambiguated scenes of the dress affects subsequent viewings, providing evidence that priors on lighting conditions can be updated by short-term experience. Two experiments were conducted with separate groups of subjects. In Experiment 1, subjects gave color matches for the original dress image immediately after viewing the disambiguating stimuli from Figure 13; in Experiment 2, subjects were never exposed to the disambiguating stimuli. In Experiment 1, two groups of subjects saw the disambiguating stimuli in one of two orders that differed depending on which disambiguating stimulus immediately preceded the color-matching task. All subjects provided verbal reports prior to viewing the disambiguating stimuli. (A) Analysis of data from Experiment 2 (online subjects: $N = 1{,}126$). The scatterplot (conventions as in Figure 1A) shows two peaks. (B) Analysis of color-matching data from Experiment 1, Order A ($N = 553$; subjects viewed the simulated warm scene—the B/K primer—directly before performing the color-matching task on the original image). *K*-means clustering returns two clusters, but there is one dominant peak. (C) Analysis of color-matching data from Experiment 1, Order B ($N = 523$; subjects viewed the simulated cool scene—the W/G primer—directly before the color-matching task on the original image). The scatterplot has two strong peaks. (D) Histograms comparing the distribution of categorical percepts recovered in Experiment 2 (control, no-primer) during the verbal task and the color-matching task. Dark-gray bars show data from the verbal reporting task; light-gray bars show the distribution of verbal reports predicted from the color matches using the category-response model (see Figure 7A; the classifier was trained on half the data and the plot shows classification for the other half). Error bars show 95% confidence intervals. Distributions do not differ. (E) Bar plots showing the distribution of categorical percepts for Experiment 1, Order A. Dark bars show verbal reports collected prior to priming; light bars show verbal labels predicted from color matches made following exposure to the B/K primer. Asterisks indicate a significant shift from W/G to B/K reporting after B/K priming (dark bar falls outside of the 95% confidence interval of the light bar). (F) Histograms comparing distributions for WG primer (Experiment 1, Order B). Photograph of the dress used with permission; copyright Cecilia Bleasdale.

with dark bars in Figure 15D), providing reassurance that we sampled a sufficient number of subjects to recover an accurate estimate of the population distribution. But compared to Figure 15A, the density plot in Figure 15B shows only one strong peak, which aligns with the color matches made by B/K subjects in the control experiment (the contour contains 80% of the data). These results show that more subjects reported B/K, and fewer reported W/G, than expected on the basis of the verbal reports that they provided. *K*-means clustering returned two optimal clusters, showing that the W/G peak was still present, albeit diminished. The bar plots (Figure 15E) quantify the shift. Figure 15C and 15F show the results for an independent set of subjects ($N = 523$) who participated in Experiment 1, Order B. Unlike with Order A, the data look similar to the control case—the density plot shows two strong peaks, and the distribution of verbal reports predicted from the color matches is not different from the distribution of verbal reports that subjects actually provided. These results show that priming subjects to see B/K influences them to see B/K, whereas priming subjects to see W/G has no effect.

What informs people's priors? Long-term exposure (chronotype)

The initial media reports, and many preliminary scientific studies, implied that a given observer's

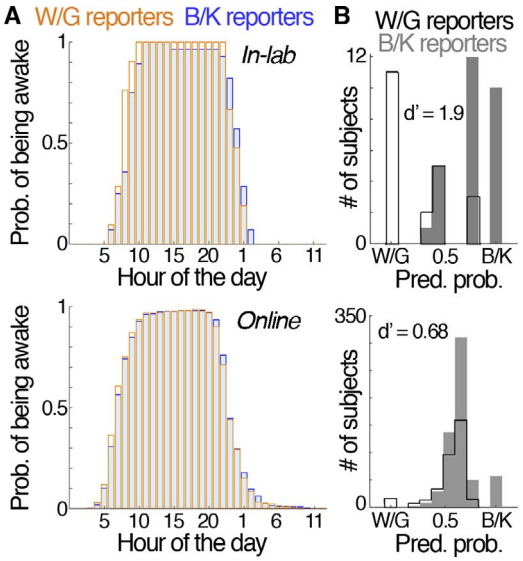

Figure 16. Relationship of chronotype to dress percept. (A) Probability of being awake at any given hour of the day as a function of dress percept (B/K reporters: blue bars; W/G reporters: gold bars; top row: in-lab, $N = 49$; bottom row: online, $N = 931$, subjects from Experiment 1). For a given hour of the day, $P(\text{awake})_{B/K}$ = (# of B/K reporters awake)/(# of B/K reporters) and $P(\text{awake})_{W/G}$ = (# of W/G reporters awake)/(# of W/G reporters). Computed from subjects' reported wake and bed hours. (B) Classification histograms for binomial logistic regression over the set of subjects' waking and sleeping hours (where each hour of the day was assigned a value of 1 or 0 and served as a predictor) to distinguish B/K from W/G reporters. In-lab: overall hit rate = 81%; B/K hit rate = 79%; W/G hit rate = 86% ($d' = 1.86$; $R^2 = 0.54$). Online: overall hit rate = 67%; B/K hit rate = 93%; W/G hit rate = 20% ($d' = 0.68$; $R^2 = 0.1$).

experience of the image was stable over time. The results we described earlier in this article support an alternative conclusion: that people can switch their perception of the image. In light of these results, we sought to revisit our first hypothesis about the origin of the individual differences in how people see the dress—that they reflect ingrained differences in the priors people hold about the spectrum of the illuminant. Most typical illuminants have chromaticities that fall along the daylight locus, which from a neutral point extends toward either a warm (orange) or a cool (blue) bias. We had hypothesized that cumulative life experience shaped the prior one holds about the spectrum of the illuminant, and that in the absence of strong cues to the illuminant, the visual system relies on these priors to discount the spectral bias in the illuminant. We speculated that the way in which the dress was seen could therefore be predicted by one's chronotype, with day larks being more like to report W/G (having a prior biased toward blue-skylight illumination) and night owls being more likely to report B/K (having a prior biased toward warm artificial light). We tested this idea by analyzing subjects' self-reported wake and bed times (hour of the day); these data were acquired at the time of our initial survey, and we queried many people who had no previous experience with the dress image. Figure 16A shows subjects' probability of being awake at any given hour of the day as a function of dress-color report. The W/G distribution is very slightly phase shifted (to the left) relative to the B/K distribution, indicating earlier wake and bed times for W/G reporters (day-lark chronotype). To test the explanatory power of chronotype, we ran a binomial regression analysis, using the set of subjects' waking and sleeping hours as predictors (each hour of the day was assigned a value of 1 or 0) to distinguish B/K from W/G reporters (Figure 16B). The regression performed above chance for the data set collected in the lab but did poorly for the online population (in-lab: correct rate = 81% overall, 79% for B/K, 86% for W/G, $d' = 1.86$, $R^2 = 0.54$; online: correct rate = 67% overall, 93% for B/K, 20% for W/G, $d' = 0.68$, $R^2 = 0.1$). We also ran a stepwise regression on the in-lab data set to determine which predictors offered the most explanatory power. In each step, a given predictive variable was considered for addition to or subtraction from the set of explanatory variables, evaluated on the basis of how well the new predictive variable improved the model. The set of explanatory variables included six specific hours of the day (7 a.m., 8 a.m., 9 a.m., 10 p.m., 11 p.m., 1 a.m.) and two interactions terms (7 a.m. with 11 p.m. and 9 a.m. with 1 a.m.).

The predictive power of the sleep–wake cycle on dress percept shown in Figure 16 is weaker than we might have expected, given the correlations reported by Wallisch (2017). There are four differences between the two studies. First, Wallisch queried subjects using links tied to news reports of the dress image; we queried subjects using surveys that were independent of news reports of the phenomenon (and many of our subjects were entirely unfamiliar with the image). Second, Wallisch asked people how they saw the dress when they first encountered it (a memory test); we asked people not only how they saw it initially but also how they saw it during the study (a perception test). Third, Wallisch asked people to self-report their chronotype (day lark versus night owl); we asked people to report the hours that they woke up and went to bed. Fourth, Wallisch surveyed about 13,000 subjects; we surveyed about 1,000.

Could these differences account for the different results in the two studies? First, a person's report of their own behavior can differ depending on the circumstances of questioning, as political polling shows: People can report favoring different candidates depending on who is asking the question. Although the news stories used to recruit participants in the Wallisch study did not themselves describe the chronotype theory, many

contemporary news stories did describe this theory (e.g., Rogers, 2015). It is conceivable that people who read the stories linked by Wallisch would have read more widely on the topic, and come across explicit descriptions of the chronotype theory. If so, could this information have biased them to report a chronotype that lined up with their dress percept? This explanation reverses the direction of causality (dress percept influencing reports of chronotype rather than chronotype influencing dress percept), and is plausible: The way we saw the dress powerfully shaped our identity. People formed camps defined by the different dress percepts that transcended race, geography, and culture; it is not unreasonable to think that this new identity influenced how we report a subjective chronotype. Second, people's memories are fallible, and not always an accurate indication of perception (exit polls differ from actual election results). Third, it is not clear that self-reports of chronotype are accurate indications of actual sleep–wake cycles. Querying the actual hour provides a direct test of the specific hypothesis that differential exposure to daylight influences the perceptual state. Querying category (day lark versus night owl) is indirect, and potentially inaccurate. One person's idea of a day lark might be very different from another's; the same category might correspond to very different amounts of daylight exposure. We are not aware of any studies showing that people would be more likely to misreport their sleep–wake cycle if queried as to the specific hour rather than the chronotype category. To the contrary, we suspect that people can accurately remember the hour they wake up and go to bed because they have several explicit markers of the time of day (alarm clocks, radio programs, TV shows).

Finally, is sample size to blame? The large sample size we used was sufficient to demonstrate categorical perceptions, and should be sufficient to uncover phenomena that are as fundamental to the way we see the dress. Nonetheless, it remains possible that our study lacked sufficient power to address the chronotype theory. One could address the question by reanalyzing the data from Wallisch to determine how much data are required, conducting an analogous power analysis to the one we present in Figure 5. Taken together, the most parsimonious conclusion of the two studies is threefold: Chronotype has a modest impact on dress percept; the way someone sees the dress can bias how they report their chronotype; and, as with other multistable images, many factors influence how the dress is perceived, and no single predictor has complete predictive power.

Do differences in where subjects report looking affect how they see the dress?

How a subject resolves the ambiguity in a multistable image such as the Necker cube can depend on where in the image a subject looks (Chastain & Burnham, 1975; Kawabata et al., 1978) or attends (Peterson & Gibson, 1991), an idea proposed by Necker himself (Long & Toppino, 2004). Toward our goal of assessing the extent to which the dress is analogous to multistable shape images, we sought to address whether subjective reports of looking behavior differed between subjects who perceived W/G and subjects who saw B/K. Subjects were asked to identify where in the image they spent most of their time looking, using a grid overlaying the image as a guide (Figure 17A, left panel). We appreciate that subjective reports of eye movements are not a good indicator of actual eye movements (Vo, Aizenman, & Wolfe, 2016; Wu & Cavanagh, 2016), and interpret these results as indications of what part of the image a subject considered most important (i.e., what part of the image they attended to). We focus our analysis on subjects who were unfamiliar with the image and who viewed the dress as either W/G or B/K, and sorted the responses on the basis of their verbal reports (Figure 17A, right panel). We draw three conclusions. First, the patterns of responses for both groups of observers are not random: Both W/G and B/K observers tend to identify regions in the center or top half of the image. This observation is consistent with other findings showing that, on average, subjects generally have a center bias (Tatler, 2007), and a top bias for visual search (Durgin, Doyle, & Egan, 2008). Second, the patterns of responses for the two groups of observers were different: B/K observers were more biased toward the upper right (shoulder) region of the dress, whereas W/G observers were more inclined to identify the center of the dress (within the fat horizontal brown stripe). Peak locations for each group are indicated with a bold line around the cell (Figure 17A). We performed a binomial regression on the reported looking locations and then did an ROC analysis to compare the differences in perceived looking behavior between the two groups (Figure 17C); the analysis shows that the differences between the groups are significant. Third, both sets of observers identified a component in the dress image that corresponds to the chromatic element that defines their perceptual state: B/K observers tend to identify a blue region, whereas W/G observers identify a brown region. These results suggest that attention to different local components (spatial frequency and color statistics) within the image play a role in determining what colors are seen in the image.

Is skin tint a sufficient cue to the illuminant?

To test the possible role of memory colors—in particular skin color—in color constancy, we deployed a variation of our disambiguation paradigm (Experiment 1; Figure 13): We asked subjects to report on the colors they saw for versions of the image that were

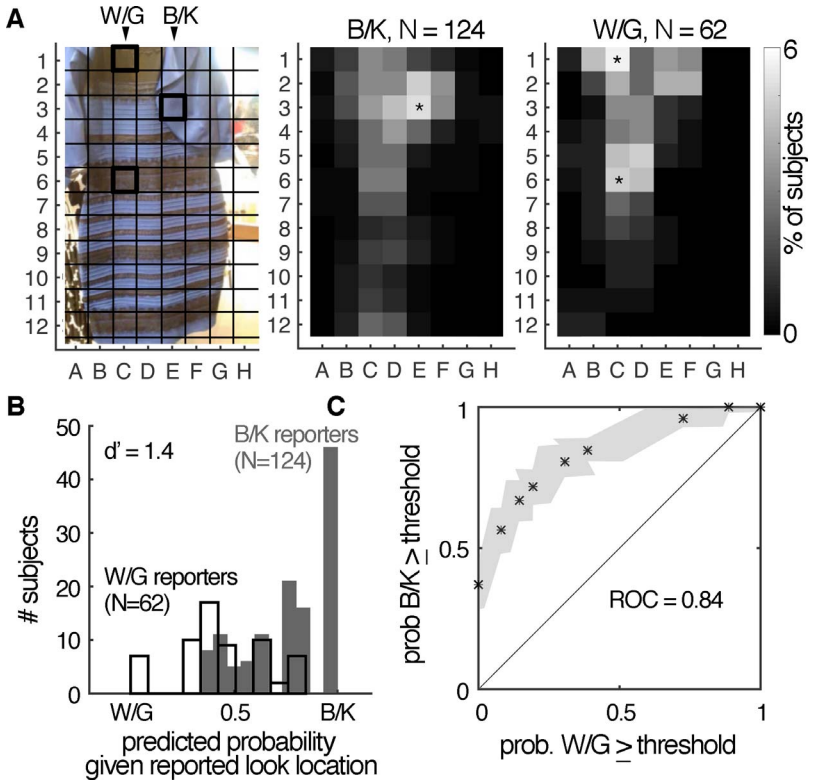

Figure 17. Self-reported looking behavior differs between B/K and W/G percievers. At the end of Experiment 1, subjects were asked to indicate the region of the image they felt they spent the most time looking at, using a grid overlaying the image as a guide. (A) Heat maps show data sorted by verbal report (left: B/K reporters; right: W/G reporters). The analysis was restricted to subjects with no prior exerience with the image (B/K = 124; W/G = 62). The maps were smoothed with a moving average of 2 × 2 grid squares. Peak locations for each group are shown with a bold outline. (B) Classification histograms from a binary logistic regression fitted using self-reported looking locations. The looking predictor variable was encoded as a binarized vector corresponding to all 96 grid locations, where reported peak look location was assigned a value of 1 and all other locations were assigned a value of 0. Histograms quanfity the probability of being B/K or W/G, given the looking reports. A test of the full model against a constant-only model was significant ($X^2 = 76$; $p = 0.008$; $R^2 = 0.33$), showing that the looking report distinguishes B/K from W/G. (C) Receiver operating characteristic (ROC) analysis confirms that the distributions in (B) are separable (ROC = 0.84; optimal threshold to assess classification accuracy is ~0.6). With a threshold of 0.6, the model hit rate was 76% for B/K and 76% for W/G ($d' = 1.44$). Photograph of the dress used with permission; copyright Cecilia Bleasdale.

manipulated to include only a low-level cue (an illuminant-biased uniform surround) or skin with an illuminant-biased tint (Figure 18). Subjects were also queried on the cue-rich stimuli from the disambiguation paradigm described in Figure 13. The RGB values of the pixels that made up the dress portion of all the disambiguation images were never modified (they retained their values from the original dress image). The experiment was carried out under controlled viewing conditions in the laboratory ($N = 53$).

When asked to rate the lighting conditions in the cue-rich scenes, subjects provided ratings consistent with the lighting conditions cued (data not shown, but are included in the analysis shown in Figure 14). Replicating our prior report (Lafer-Sousa et al., 2015) and the findings in Figure 13, and consistent with color-constancy predictions, most subjects' percepts—regardless of how subjects initially perceived the dress's colors—conformed to the percept predicted by the color of the illuminant cued. The distribution of subject percepts changed dramatically when we changed the context from warm to cool ($p < 0.001$, paired t test): When cool light was cued, most subjects reported W/G (Figure 18A, left panel); when warm light was cued, most subjects reported B/K (Figure 18A, right panel). To test whether a low-level sensory mechanism like receptor adaptation or local color contrast (long argued to play a powerful role in color constancy; Brainard & Wandell, 1992; Hurlbert & Wolf, 2004; Land, 1986) could disambiguate the dress's colors, we superimposed the isolated dress on uniform fields matched to the mean chromaticity of the cue-rich scenes. If low-level sensory adaptation or local color contrast is sufficient to achieve color constancy, surrounding the dress by a yellowish field should induce a uniform B/K percept, while surrounding it by a bluish

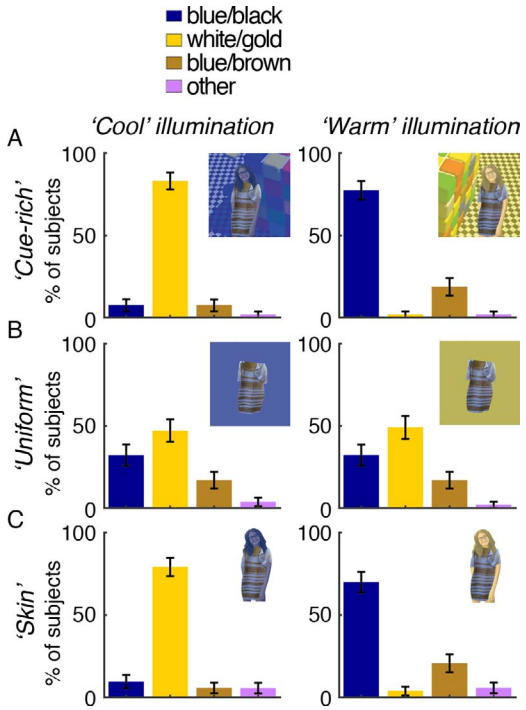

Figure 18. Behavioral evidence that people can achieve color constancy by using human skin as a reference to ascertain the spectral bias of the light source. Histograms showing subjects' reports of the dress's color when it was embedded in simulated contexts designed to convey cool illumination (left column) or warm illumination (right): (A) cue-rich scenes; (B) uniform illuminant-biased chromatic backgrounds; (C) skin with illuminant-biased tints, in isolation. The color of the bar corresponds to the verbal reports (key at top). Insets show stimulus thumbnail (see Supplementary Image Appendix for full-size reproductions). Error bars were calculated using bootstrapping (10,000 bootstrapped samples). $N = 53$ subjects. See Table 2 for McNemar's chi-square tests comparing goodness of constancy achieved across stimulus conditions. Photograph of the dress used with permission; copyright Cecilia Bleasdale.

Test comparison	X^2	p
Cue-rich scene (warm) vs. uniform background (warm)	22.04	$\mathbf{3 \times 10^{-6}}$
Cue-rich scene (cool) vs. uniform background (cool)	15.43	$\mathbf{8 \times 10^{-5}}$
Skin tint (warm) vs. uniform background (warm)	15.04	$\mathbf{1 \times 10^{-4}}$
Skin tint (cool) vs. uniform background (cool)	15.06	$\mathbf{1 \times 10^{-4}}$
Skin tint (warm) vs. cue-rich scene (warm)	2.25	0.1
Skin tint (cool) vs. cue-rich scene (cool)	0.25	0.6

Table 2. McNemar's chi-square tests comparing goodness of constancy achieved in different stimulus conditions (see Figure 18). McNemar's test is a within-subject z test of equality of proportions for repeated measures. Each test compares the proportion of subjects that did or did not conform to the percept cued in stimulus condition X versus the proportion of subjects that did or did not conform to the percept cued in stimulus condition Y.

field should induce a W/G percept. Contrary to this prediction, when the dress was surrounded by the uniform chromatic backgrounds subjects' percepts did not conform to the percept predicted by the color of the background; the distribution of subject percepts was the same for both backgrounds ($p = 0.71$, paired t test; Figure 18B). The cue-rich scenes provided significantly better color constancy than the uniform surrounds (summary statistics provided in Table 2). Taken together, these results suggest that a strictly low-level mechanism is insufficient to elicit a stable color percept of the dress.

To test whether skin chromaticity is by itself a sufficient cue to achieve good color constancy, we presented the dress superimposed on the woman on a white background and tinted her skin according to the spectral bias of the illuminants simulated in the cue-rich scenes. Consistent with the hypothesized role of skin memory color in discounting the illuminant from a scene, we found that subjects' color percepts were predicted by the color bias of the skin tint. The distribution of subject percepts changed dramatically between the warm and cool conditions ($p = 0.002$, paired t test): When the tint cued cool light, most subjects reported W/G (Figure 18C, left panel); when the tint cued warm light, most subjects reported B/K (Figure 18C, right panel). The presence of tinted skin alone was significantly more effective than the uniform illuminant-biased background in achieving good color constancy, and was as effective as the cue-rich scene (Table 2). These results suggest that people can use skin chromaticity to recover information about the spectral composition of the illuminant to achieve color constancy.

Discussion

This study addresses the extent to which the dress image is multistable. For an image to be multistable, it is widely assumed that it must appear to change state (flip or reverse) on a short timescale (the Necker cube is typically seen to pop out, and then a few seconds later to recede). But a review of the literature (discussed in the following) suggests that rapid flipping is not a necessary feature of stimuli that have been considered multistable. Rather, the key properties of multistable stimuli are that they have more than one plausible percept; that the alternative percepts are mutually

exclusive; and that switches between the alternative percepts can occur within single individuals. By these criteria, the results presented here provide the first evidence for a multistable color image. Quantitative analysis of color-matching data from a large, diverse population shows that the dress image is experienced categorically as either blue and black (B/K) or white and gold (W/G). A power analysis shows that several hundred observers are needed to uncover the categorical nature of the population distribution. The two perceptual states were distinguished not only by differences in lightness, as suggested previously (Gegenfurtner et al., 2015), but also by differences in hue. This discovery is important: A multistable image with states defined only by differences in lightness might not, arguably, be considered a multistable *color* image. Classifiers trained on the color-matching data performed very well: Given your color-matching data, the algorithm can very likely report how you describe the dress. These results not only support the idea that the dress is multistable but also provide an important example of subjective color experiences being predicted from objective colorimetric data.

Color matches made by the population of B/K observers covered a larger region of color space than those made by W/G observers. This result may reflect the asymmetry in blue/yellow perception: Changes in yellowness are more detectable than changes in blueness (Winkler, Spillmann, Werner, & Webster, 2015). The asymmetry may also reflect the lower communicative efficiency of cool colors compared to warm colors (Gibson et al., 2017). Color matches to the bottom region of the dress gave rise to three clusters instead of two. One author (BRC) initially saw the dress as blue/brown (B/B), and so we were initially inclined to interpret the third cluster as demarking a stable B/B category. But the cluster peak falls outside of the color matches made by observers who report B/B. Seeing the dress as B/B appears to be a transition state. Our classifier accurately classified only half of B/B reporters, and only with relatively low probability. Among all groups of subjects, those reporting B/B (and "other") also reported the highest rates of changing their minds about the color of the dress. Rather than identifying a stable B/B category, the additional cluster recovered by the cluster analysis subdivides the people who see the dress as B/K into separate dark-blue and light-blue reporters. This subdivision of blue is reminiscent of some languages that carve blue into distinct light and dark categories (Winawer et al., 2007).

The results we obtained in the lab were generally consistent with those obtained online, suggesting that the neural operations that give rise to different perceptions of the dress are largely invariant to low-level stimulus features such as mean luminance and white point, which vary among different displays. These results contribute to a growing body of work that exploits the greater power and participant diversity afforded by online experiments (Wilmer et al., 2012).

#TheDress: Evidence that it is multistable

When the image first went viral on the Internet, many people stated that they had a stable perception of the dress and were surprised that others saw it differently (Rogers, 2015). The apparent lack of reversibility is evidence against the notion that the image is analogous to multistable shape images such as the Necker cube. But is a person's experience of the dress really fixed? We found that about 50% of subjects reported a reversal prior to participating in our experiment; about 12% reported that their experience of reversals was frequent. We did not ask subjects if their perception of the dress flipped while they performed our experiment, but several subjects tested in the lab spontaneously reported that it did. Perceptual reversals of the dress are probably not as rare as initially suggested in the popular press. The general assumption that one's perception of the image is fixed may partly be accounted for by the fact that people do not look at the image for very long. Given our own experience, we suspect that requiring subjects to maintain engagement with the image for longer periods of time would promote reversals; extended, continuous exposure promotes reversals for other multistable images (Leopold, Wilke, Maier, & Logothetis, 2002). Consistent with the role of top-down factors in determining how multistable shape images are seen (Scocchia et al., 2014), we found that knowledge of the dress's color in real life increased the likelihood that subjects experienced a reversal, as did having professional illustration and graphic-design experience. Over time the population has gravitated toward seeing the dress as B/K, its colors in real life. That peoples' perception has drifted toward the true colors of the dress is reminiscent of some other ambiguous images, such as the Dalmatian image (Gregory, 1970) and Mooney faces, where knowledge of what is being depicted has a profound (and stable) impact on how the ambiguous image is subsequently seen (Lupyan, 2017). But knowledge of the dress's true colors is not by itself decisive; many people who know its true colors maintain that the image appears W/G, suggesting the dress is more similar to a multistable image than a simple ambiguous one.

Although we found some evidence that people can change their mind about the color of the dress, most observers said that the image did not flip spontaneously or frequently, which appears to contradict the idea that the dress is multistable. But is frequent perceptual flipping a necessary feature of the class of stimuli that

should be grouped together as being multistable? Perhaps not. The reversal frequency of multistable images varies from individual to individual and stimulus to stimulus; the ease of reversibility depends on (among other factors) ambiguity type and prior knowledge. For example, content reversals and figure–ground reversals occur with higher frequency than perspective reversals (Kosegarten & Kose, 2014). Some multistable images, such as the rotating dancer, are notoriously "sticky" (some people never spontaneously see the dancer change direction). Moreover, subjects rarely report spontaneous reversals unless they are told that the images can be seen in different states, even for the Necker cube (Rock & Mitchener, 1992). And even knowledge of multistability is not always sufficient to induce reversals (Kosegarten & Kose, 2014). That most people's experience of the dress image is relatively stable is not sufficient to reject the idea that the image is multistable. The stickiness of the dress image may simply indicate that the visual apparatus prioritizes stable and consistent color perception, which may explain why so few examples of multistability in color have been discovered.

The perception of multistable images is affected not only by top-down factors such as prior knowledge but also by bottom-up factors (Long & Toppino, 2004) such as the local image statistics at the point of gaze (Chastain & Burnham, 1975; Kawabata et al., 1978; Kawabata & Mori, 1992). Reducing differences in texture between the dress and the surround (by blurring or shrinking the image) increased reports of B/K (Lafer-Sousa et al., 2015). When asked to identify the region they looked at most, subjects who reported B/K selected a blue region, whereas subjects who reported W/G picked a gold region. Thus all subjects tended to discount as achromatic the component that they thought they were not looking at. Where one looks will bias the interpretation of color statistics across a scene (Toscani, Valsecchi, & Gegenfurtner, 2015) and influence color induction (Brenner, Granzier, & Smeets, 2007; Hansen & Gegenfurtner, 2005), providing an explanation for the looking behavior: B/K reporters tended to focus on the right edge of the dress, where color judgments would reflect stronger induction by the warm background. W/G reporters often identified regions in the center of the image, away from the warm background. Curiously, the most common region identified across all subjects tended to cover a large area of uniform color (B/K: the broad blue shoulder; W/G: the wide dark stripe), perhaps reflecting the importance of low spatial frequencies in color constancy (Dixon & Shapiro, 2017). Although these analyses suggest that bottom-up factors shape how we see the dress, it is possible that where subjects say they look is caused by how they see the dress's colors. Taken together, the evidence presented here suggests that color images can be bistable. Documentation of another such image will be needed to prove the case. One possible case is the pink/white versus gray/teal shoe image, which shares many similar features to the dress image, and importantly, many people have reported spontaneously switching their perception of the image (The Guardian, 2017; Huffington Post, 2017). Both the shoe image and the dress image are comprised of an achromatic protion and the chromatic portion; and both with ambiguous cues to the lighting.

#TheDress: Assumptions about the illumination that are updated over short and long timescales explain the individual differences

In our preliminary report, we argued that the image is consistent with two different interpretations: that the dress is in shadow, illuminated by bluish light (in which case people see the dress as W/G), or that it is illuminated by a warm light (in which case people see it as B/K). Here we tested this hypothesis directly by analyzing the results of the subjects' reports of the lighting conditions. Indeed, the largest factor that determined how the dress was seen was how subjects interpreted the illumination: Subjects who thought the dress was in shadow were much more likely to see it as W/G. Our results show that the inferences about the lighting correspond to whether subjects segment the scene into one or two frameworks (Zdravkovic, Economou, & Gilchrist, 2012): On the one hand, B/K reporters make similar judgments about the light illuminating the background and the light illuminating the foreground (one framework: bright, warm direct/global illumination); on the other hand, W/G reporters make dissimilar judgments about the background and foreground lighting (two frameworks: bright, warm background light, but dim/cool foreground light, consistent with cast shadow).

The two different interpretations of the dress's colors may be enabled because the chromaticities of the pixels within it are restricted to the colors of the daylight axis (Lafer-Sousa et al., 2015): When viewed in isolation, all the pixels are either bluish or orangish. There is no other chromatic information available for the visual system to resolve the lighting conditions. We argued that people must rely not only on other cues to sort out the lighting conditions but also on priors about what they assumed the lighting conditions to be. These priors are presumably established over both a short time frame (what time of day is it now? What illuminant were you recently experiencing?) and a longer time frame (what sort of light are you most often exposed to?). We tested for priors acquired over both timescales. Priming subjects with an unambiguous version of the image biased subjects to make color matches consistent with that

unambiguous version, regardless of how they said they saw the image initially. These results provide evidence that subjects flipped their perception during the experiment, and support the idea that the brain updates its illumination priors as soon as reliable information about the lighting is obtained. But curiously, we saw these effects only in one direction: Priming with a warm illuminant predisposed people to make color matches consistent with B/K, but priming with a cool illuminant had little impact. We do not have an explanation for the asymmetry, although it is consistent with three other observations showing that W/G is less stable over time: The B/K report was increased by (1) reduced image size, (2) any prior exposure to the image, and (3) knowledge of the color of the dress in the real world.

In the initial media reports of the dress, we outlined a chronotype theory to account for the individual differences (Lafer-Sousa et al., 2015; Rogers, 2015). According to this theory, day larks see the dress as W/G while night-owls see it as B/K. The idea was that day larks are more exposed to a blue spectral bias in the illuminant, and so are more likely to discount the blue component of the dress and see the dress as W/G; night owls are more exposed to incandescent light with an orange spectral bias, and are therefore more likely to discount the warm component in the image and see it as B/K. The explanatory power of chronotype in our study was larger for the in-lab data than the online data. All in-lab subjects in our study were residents of the Boston area, and so were matched for regional light cycle and atmospheric conditions. The participants in the online experiments were distributed across the globe, which may account for the noisier results obtained from this population. Nonetheless, the predictive power of sleep–wake cycle on dress percept that we report here is considerably weaker than we would have expected, given the correlations reported by Wallisch (2017). One salient difference between the studies is the way in which subjects were recruited: Wallisch recruited subjects through online links attached to news reports of the image. While we acknowledge the appeal of the idea that different chronotypes underlie the individual differences in perception of the dress, it is possible that the causal link is the reverse. People formed camps defined by the different dress percepts, which transcended race, geography, and culture; it is not unreasonable to think that this new identity influenced how a person reported their chronotype in the study by Wallisch. But as we note earlier, there are also differences in the number of subjects used in our study and Wallisch's, which might give rise to differences in experimental power. Taken together, the most parsimonious conclusion of the two studies is threefold: Chronotype has a modest impact on dress percept (that might be confounded by age and gender); the way someone sees the dress can bias how they report their chronotype; and, as with other multistable images, many factors influence how the dress is perceived, and no single predictor has complete predictive power.

#TheDress: A tool for understanding color

Finally, we showcase the dress image as a tool to test the role of memory in color constancy. An estimate of the spectrum of the illuminant can theoretically be achieved by performing gamut correlation between an observed object and a memory gamut for that object (Ling & Hurlbert, 2008). In particular, skin has been proposed as a potential cue for color constancy, due to its stable statistics (Bianco & Schettini, 2012; Crichton et al., 2012; Hurlbert, 2007). Using digitally manipulated versions of the dress, we show that a color tint applied to the skin of a model made to look like she is wearing the dress was sufficient for observers to infer the spectral bias of the illuminant and achieve a predictable percept of the dress's colors. Placing the dress on a uniform colored background that matched the spectral bias of a light source was insufficient for subjects to disambiguate the colors of the dress. These results provide a striking demonstration that color constancy exploits more than local color contrast. The importance of skin for color constancy might have been predicted not only by the ubiquity of skin in visual experience but also by the observation that humans have precise memory colors and preferences for skin, and show little tolerance for color deviations in reproductions of skin (Bartleson & Bray, 1962; Chauhan, Xiao, Yates, & Wuerger, 2015; Hunt, Pitt, & Winter, 1974; Sanders, 1959; Smet, Ryckaert, Pointer, Deconinck, & Hanselaer, 2011; Yendrikhovskij, Blommaert, & Ridder, 1999). Moreover, people are most sensitive to skin color adjustments that roughly align with the spectral bias of natural illuminants (the daylight axis; Wuerger, Chauhan, Sohaib, Yates, & Xiao, 2016). We only tested illumination cues with Caucasian skin, assessed mostly by Caucasian participants. But given computational analysis of the statistics of different skin types (Crichton et al., 2012), we expect that our results would generalize across races. We carry skin with us everywhere, so it is perhaps not surprising that the visual system uses it like a color-balance card in photography, to correct the bias in the illuminant.

Keywords: color categorization, color constancy, bistable illusion

Acknowledgments

This work was supported by the Intramural Research Program of the National Eye Institute and

grants from the National Institutes of Health (R01 EY023322) and the National Science Foundation (1353571). We are grateful to Katherine Hermann for providing the custom data-acquisition code and for critical contributions to the design of the study. We also thank Edward Adelson, Nancy Kanwisher, James DiCarlo, and members of the Laboratory of Sensorimotor Research at the National Eye Institute for valuable feedback.

Commercial relationships: none.
Corresponding author: Bevil R. Conway.
Email: bevil@nih.gov.
Address: Laboratory of Sensorimotor Research, National Eye Institute, and National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA.

Footnote

[1] http://www.mathworks.com/matlabcentral/fileexchange/8577-scatplot

References

Bartleson, C. J., & Bray, C. P. (1962). On the preferred reproduction of flesh, blue-sky, and green-grass colors. *Photographic Science and Engineering*, *6*, 19–25.

Bianco, S., & Schettini, R. (2012). Color constancy using faces. Presented at the IEEE Conference on Computer Vision and Pattern Recognition, June 2012, Providence, RI, USA.

Bloj, M. G., Kersten, D., & Hurlbert, A. C. (1999). Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature*, *402*(6764), 877–879.

Brainard, D. H., & Hurlbert, A. C. (2015). Colour vision: Understanding #theDress. *Current Biology*, *25*(13), R551–R554.

Brainard, D. H., Longere, P., Delahunt, P. B., Freeman, W. T., Kraft, J. M., & Xiao, B. (2006). Bayesian model of human color constancy. *Journal of Vision*, *6*(11):10, 1267–1281, doi:10.1167/6.11.10. [PubMed] [Article]

Brainard, D. H., & Wandell, B. A. (1986). Analysis of the retinex theory of color vision. *Journal of the Optical Society of America A*, *3*(10), 1651–1661.

Brainard, D. H., & Wandell, B. A. (1992). Asymmetric color matching: How color appearance depends on the illuminant. *Journal of the Optical Society of America A*, *9*(9), 1433–1448.

Brenner, E., Granzier, J. J., & Smeets, J. B. (2007). Perceiving colour at a glimpse: The relevance of where one fixates. *Vision Research*, *47*(19), 2557–2568.

Chastain, G., & Burnham, C. A. (1975). The first glimpse determines the perception of an ambiguous figure. *Perception & Psychophysics*, *17*(3), 221–224.

Chauhan, T., Xiao, K., Yates, J., & Wuerger, S. (2015). Estimating discrimination ellipsoids for skin images. *Journal of Vision*, *15*(12):820, doi:10.1167/15.12.820. [Abstract]

Chetverikov, A., & Ivanchei, I. (2016). Seeing "the Dress" in the right light: Perceived colors and inferred light sources. *Perception*, *45*(8), 910–930.

Chichilnisky, E. J., & Wandell, B. A. (1995). Photoreceptor sensitivity changes explain color appearance shifts induced by large uniform backgrounds in dichoptic matching. *Vision Research*, *35*(2), 239–254.

Conway, B. R. (2015). Why do we care about the colour of the dress? *The Guardian*. Retrieved from www.theguardian.com/commentisfree/2015/feb/27/colour-dress-optical-illusion-social-media

Conway, B. R. (2016). Processing. In C. A. Jones, D. Mather, & R. Uchill (Eds.), *Experience: Culture, cognition, and the common sense* (pp. 86–109). Cambridge, MA: MIT Press and the MIT Center for Art, Science & Technology.

Crichton, S., Pichat, J., Mackiewicz, M., Tian, G., & Hurlbert, A. C. (2012, Jan). *Skin chromaticity gamuts for illumination recovery*. Paper presented at the Society for Imaging Science and Technology Conference on Color in Graphics, Imaging, and Vision, Amsterdam, the Netherlands.

Dixon, E. L., & Shapiro, A. G. (2017). Spatial filtering, color constancy, and the color-changing dress. *Journal of Vision*, *17*(3):7, 1–15, doi:10.1167/17.3.7. [PubMed] [Article]

Drissi Daoudi, L., Doerig, A., Parkosadze, K., Kunchulia, M., & Herzog, M. H. (2017). The role of one-shot learning in #TheDress. *Journal of Vision*, *17*(3):15, 1–7, doi:10.1167/17.3.15. [PubMed] [Article]

Durgin, F. H., Doyle, E., & Egan, L. (2008). Upper-left gaze bias reveals competing search strategies in a reverse Stroop task. *Acta Psychologica*, *127*(2), 428–448.

Dutour, E.-F. (1760). Discussion d'une question d'optique. *l'Académie des Sciences: Mémoires de*

Mathématique et de Physique Présentés par Divers Savants, *3*, 514–530.

D'Zmura, M., & Lennie, P. (1986). Mechanisms of color constancy. *Journal of the Optical Society of America A*, *3*(10), 1662–1672.

Ellis, S. R., & Stark, L. (1978). Eye movements during the viewing of Necker cubes. *Perception*, *7*(5), 575–581.

Fisher, G. H. (1967). Preparation of ambiguous stimulus materials. *Perception & Psychophysics*, *2*, 421–422.

Foster, D. H., Nascimento, S. M. C. (1994). Relational colour constancy from invariant cone-excitation ratios. *Proceedings of the Royal Society B: Biological Sciences*, *257*, 115–121.

Gegenfurtner, K. R., Bloj, M., & Toscani, M. (2015). The many colours of "the dress." *Current Biology*, *25*(13), R543–R544.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasignam, S., et al. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences USA*, under review.

Gregory, R. (1970). *The intelligent eye*. New York: McGraw-Hill.

Guardian, The. (2017). *Shoe colour question could put 2015 dress debate in the shade*. Retrieved from https://www.theguardian.com/technology/2017/oct/12/shoe-colour-question-could-put-2015-dress-debate-in-the-shade

Hansen, T., & Gegenfurtner, K. R. (2005). Classification images for chromatic signal detection. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *22*(10), 2081–2089.

Hedrich, M., Bloj, M., & Ruppertsberg, A. I. (2009). Color constancy improves for real 3D objects. *Journal of Vision*, *9*(4):16, 1–16, doi:10.1167/9.4.16. [PubMed] [Article]

Huffington Post, The. (2017). *Much like 'The Dress,' people can't decide what color these shoes are*. Retrieved from https://www.huffingtonpost.com/entry/people-cant-shoes-grey-teal-mint-pink-white_us_59e0bcc7e4b04d1d51811f2b

Hunt, R. W. G., Pitt, I. T., & Winter, L. M. (1974). The preferred reproduction of blue sky, green grass and Caucasian skin in colour photography. *Journal of Photographic Science*, *22*, 144–149.

Hurlbert, A. (2007). Colour constancy. *Current Biology*, *17*(21), R906–R907.

Hurlbert, A., & Wolf, K. (2004). Color contrast: A contributory mechanism to color constancy. *Progress in Brain Research*, *144*, 147–160.

Ishihara, S. (1977). *Tests for colour blindness*. Tokyo: Kanehara Shuppen Company.

Kawabata, N., & Mori, T. (1992). Disambiguating ambiguous figures by a model of selective attention. *Biological Cybernetics*, *67*(5), 417–425.

Kawabata, N., Yamagami, K., & Noaki, M. (1978). Visual fixation points and depth perception. *Vision Research*, *18*(7), 853–854.

Khang, B. G., & Zaidi, Q. (2002). Cues and strategies for color constancy: Perceptual scission, image junctions and transformational color matching. *Vision Research*, *42*(2), 211–226.

Klink, P. C., van Wezel, R. J., & van Ee, R. (2012). United we sense, divided we fail: Context-driven perception of ambiguous visual stimuli. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1591), 932–941.

Kosegarten, J., & Kose, G. (2014). Seeing reversals in ambiguous images: To know or not to know? *Perceptual and Motor Skills: Perception*, *119*(1), 228–236.

Kraft, J. M., & Brainard, D. H. (1999). Mechanisms of color constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences, USA*, *96*(1), 307–312.

Lafer-Sousa, R., Hermann, K. L., & Conway, B. R. (2015). Striking individual differences in color perception uncovered by "the dress" photograph. *Current Biology*, *25*(13), R545–R546.

Land, E. H. (1986). Recent advances in retinex theory. *Vision Research*, *26*(1), 7–21.

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences*, *3*(7), 254–264.

Leopold, D. A., Wilke, M., Maier, A., & Logothetis, N. K. (2002). Stable perception of visually ambiguous patterns. *Nature Neuroscience*, *5*(6), 605–609.

Ling, Y., & Hurlbert, A. (2008). Role of color memory in successive color constancy. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *25*(6), 1215–1226.

Long, G. M., & Toppino, T. C. (2004). Enduring interest in perceptual ambiguity: Alternating views of reversible figures. *Psychological Bulletin*, *130*(5), 748–768.

Lotto, R. B., & Purves, D. (2002). The empirical basis of colour perception. *Consciousness and Cognition*, *11*(4), 609–629.

Lupyan, G. (2017). Changing what you see by changing what you know: The role of attention. *Frontiers in Psychology*, *8*, 553.

Peterson, M. A., & Gibson, B. S. (1991). Directing

spatial attention within an object: Altering the functional equivalence of shape descriptions. *Journal of Experimental Psychology*, *17*(1), 170–182.

Podvigina, D. N., & Chernigovskaya, T. V. (2015). Top-down influences to multistable perception: Evidence from temporal dynamics. *International Scholarly and Scientific Research and Innovation*, *9*(11), 3849–3852.

Rock, I., & Mitchener, K. (1992). Further evidence of failure of reversal of ambiguous figures by uninformed subjects. *Perception*, *21*(1), 39–45.

Rogers, A. (2015). The science of why no one agrees on the color of this dress. *Wired*. Retrieved from www.wired.com/2015/02/science-one-agrees-color-dress/

Sanders, C. L. (1959). Color preferences for natural objects. *Illuminating Engineering*, *54*, 452–456.

Schwartz, J. L., Grimault, N., Hupe, J. M., Moore, B. C., & Pressnitzer, D. (2012). Multistability in perception: Binding sensory modalities, an overview. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1591), 896–905.

Scocchia, L., Valsecchi, M., & Triesch, J. (2014). Top-down influences on ambiguous perception: The role of stable and transient states of the observer. *Frontiers in Human Neuroscience*, *8*, 979.

Smet, K. A. G., Ryckaert, W. R., Pointer, M. R., Deconinck, G., & Hanselaer, P. (2011). Colour appearance rating of familiar real objects. *Color Research and Application*, *36*, 192–200.

Stiles, W. S. (1959). Color vision: The approach through increment threshold sensitivity. *Proceedings of the National Academy of Sciences, USA*, *45*, 100–114.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14):4, 1–17, doi: 10.1167/7.14.4. [PubMed] [Article]

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society B*, *63*, 411–423.

Toscani, M., Gegenfurtner, K. R., & Doerschner, K. (2017). Differences in illumination estimation in #thedress. *Journal of Vision*, *17*(1):22, 1–14, doi:10.1167/17.1.22. [PubMed] [Article]

Toscani, M., Valsecchi, M., & Gegenfurtner, K. R. (2015). Effect of fixation positions on perception of lightness. In B. E. Rogowitz, T. N. Pappas, & H. de Ridder (Eds.), *Proceedings of SPIE 9394, human vision and electronic imaging* (93940R). Bellingham, WA: SPIE, doi:10.1117/12.2175673.

Vo, M. L., Aizenman, A. M., & Wolfe, J. M. (2016). You think you know where you looked? You better look again. *Journal of Experimental Psychology: Human Perception & Performance*, *42*(10), 1477–1481.

von Kries, J. (1878). Beitrag zur Physiologie der Gesichtsempfindugen. In D. L. MacAdam (Ed.), *Sources of color science* (1st ed., pp. 101–108). Cambridge, MA: MIT Press.

Wallisch, P. (2017). Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain: "The dress." *Journal of Vision*, *17*(4):5, 1–14, doi:10.1167/17.4.5. [PubMed] [Article]

Webster, M. A., & Mollon, J. D. (1995). Colour constancy influenced by contrast adaptation. *Nature*, *373*(6516), 694–698.

Webster, M. A., & Mollon, J. D. (1997). Adaptation and the color statistics of natural images. *Vision Research*, *37*(23), 3283–3298.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology*, *29*(5–6), 360–392.

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, USA*, *104*(19), 7780–7785.

Winkler, A. D., Spillmann, L., Werner, J. S., & Webster, M. A. (2015). Asymmetries in blue-yellow color perception and in the color of "the dress." *Current Biology*, *25*(13), R547–R548.

Witzel, C., Racey, C., & O'Regan, J. K. (2017). The most reasonable explanation of "the dress": Implicit assumptions about illumination. *Journal of Vision*, *17*(2):1, 1–19, doi:10.1167/17.2.1. [PubMed] [Article]

Wu, D. A., & Cavanagh, P. (2016). Where are you looking? Pseudogaze in afterimages. *Journal of Vision*, *16*(5):6, 1–10, doi:10.1167/16.5.6. [PubMed] [Article]

Wuerger, S., Chauhan, T., Sohaib, A., Yates, J., & Xiao, K. (2016). The sensitivity of the human visual system to subtle skin tone changes. *Journal of Vision*, *16*(12):643, doi:10.1167/16.12.643. [Abstract]

Yendrikhovskij, S. N., Blommaert, F. J. J., & Ridder, H. (1999). Color reproduction and the naturalness constraint. *Color Research and Application*, *24*, 54–67.

Zdravkovic, S., Economou, E., & Gilchrist, A. (2012). Grouping illumination frameworks. *Journal of Experimental Psychology: Human Perception & Performance*, *38*(3), 776–784.