Vision Research 141 (2017) 76–94

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

The dress and individual differences in the perception of surface properties

Christoph Witzel [a,*], J. Kevin O'Regan [b], Sabrina Hansmann-Roth [c,d]

[a] Allgemeine Psychologie, Justus-Liebig-Universität Gießen, Gießen, Germany
[b] Laboratoire Psychologie de la Perception (UMR 8242), Université Paris Descartes, Paris, France
[c] Laboratoire des Systèmes Perceptifs (UMR 8248 CNRS), Ecole Normale Supérieure, PSL Research University, Paris, France
[d] Departement d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, Paris, France

ARTICLE INFO

Article history:
Received 28 October 2016
Received in revised form 30 June 2017
Accepted 5 July 2017
Available online 1 September 2017

Number of Reviews = 3

Keywords:
Individual differences
#The Dress
Colour constancy
Material properties

ABSTRACT

This study investigates systematic individual differences in the way observers perceive different kinds of surface properties and their relationship to the dress, which shows striking individual differences in colour perception. We tested whether these individual differences have a common source, namely differences in perceptual strategies according to which observers attribute features in two-dimensional images to surfaces or to their illumination. First, we reanalysed data from two previous experiments on the dress and colour constancy. The comparison of the two experiments revealed that the colour perception of the dress is strongly related to individual differences in colour constancy. Second, two online surveys measured individual differences in the perception of colour-ambiguous images including the dress, in colour constancy, in gloss perception, in the subjective grey-point, in colour naming, and in the perception of an image with ambiguous shading. The results of the surveys replicated and extended previous findings according to which individual differences in the colour perception of the dress are due to implicit assumptions about the illumination. However, results also showed that the individual differences for other phenomena were independent of the dress and of each other. Overall, these results suggest that the striking individual differences in dress colour perception are due to individual differences in the interpretation of illumination cues to achieve colour constancy. At the same time, they undermine the idea of an overall perceptual strategy that encompasses other phenomena more generally related to the interpretation of illumination and surface properties.

1. Introduction

The photo of a dress (Swiked, 2015) has highlighted the importance of individual differences in perception because it revealed striking individual differences in colour perception: Many observers saw the dress blue and black, while almost all the others saw it white and gold (Bach, 2015; Brainard & Hurlbert, 2015; Gegenfurtner, Bloj, & Toscani, 2015; Lafer-Sousa, Hermann, & Conway, 2015; Macknik, Martinez-Conde, & Conway, 2015; Swiked, 2015). It has been shown that these striking individual differences are related to observers' implicit assumptions about the illumination of the scene on the photo (Chetverikov & Ivanchei, 2016; Hesslinger & Carbon, 2016; Toscani, Dörschner, & Gegenfurtner, 2016; Wallisch, 2017; Witzel, Racey, & O'Regan, 2017). However, it is not yet clear why different observers interpret the photo differently.

Some have argued that the individual differences in the perception of the dress are due to hard-wired, sensory differences in perceptual processing. This view is supported by evidence that twins tend to see the colours similarly (Mahroo et al., 2017), and that white-gold seers tend to have larger pupil sizes (Vemuri, Bisla, Mulpuru, & Varadharajan, 2016) and higher macular pigment optical density (Rabin, Houser, Talbert, & Patel, 2016). Moreover, the observation that the perception of the dress is related to gender and age might also be taken as evidence for hard-wired determinants, such as age-related changes in the eye (Lafer-Sousa et al., 2015; Mahroo et al., 2017; Moccia et al., 2016; Wallisch, 2017).

However, the observation that a few observers can switch between different perceptions speaks against a hard-wired origin of the individual differences (Bach, 2015; Lafer-Sousa et al., 2015; Witzel, 2015). Moreover, the ambiguity in the perception of the dress is rather specific to that photograph. Hard-wired dif-

* Corresponding author.
E-mail address: cwitzel@daad-alumni.de (C. Witzel).

http://dx.doi.org/10.1016/j.visres.2017.07.015
0042-6989/

ferences cannot explain why the striking individual differences occur for this particular photo of the dress, but not in many other situations involving colour perception in everyday life (Bach, 2015; Witzel, 2015).

It has also been proposed that colour naming might play a role for the individual differences in the description of the dress (Bach, 2015). There are substantial individual differences in colour naming, even when using the most basic colour terms, such as yellow, green, blue, or purple (Lindsey & Brown, 2009; Olkkonen, Witzel, Hansen, & Gegenfurtner, 2010; Webster & Kay, 2007; Witzel & Gegenfurtner, 2013). A link between colour naming and reported dress colours is to some extent supported by the observation that reported dress colours are related to individual differences in blue ratings along a white-blue continuum (Hesslinger & Carbon, 2016). However, differences in category boundaries cannot account for the complete phenomenon because the black and gold categories are not adjacent and hence there is no direct boundary between them (Witzel, 2015). More importantly, it has been shown that the individual differences in reported dress colours constitute a continuous perceptual, rather than categorical linguistic phenomenon (Gegenfurtner et al., 2015; Lafer-Sousa et al., 2015; Witzel, Racey, & O'Regan, 2017). In sum, even if individual differences in hard-wired colour processing and in colour naming may have some influence on the perception of the dress colours they cannot account for the strong and surprising effects that are particular to the dress phenomenon.

An alternative view proposes that the perception of the dress colours in the photo is a special case of colour constancy (Bach, 2015; Brainard & Hurlbert, 2015; Macknik et al., 2015; Witzel, 2015; Witzel et al., 2017). Colour constancy allows observers to identify the colour of an object's surface despite the fact that changes in illumination can create dramatic changes in the sensory colour signal received at the retina (as quantified by colorimetric Tristimulus Values and cone excitations). According to this view, the illumination in the photo is ambiguous and observers unconsciously infer the illumination of the real three-dimensional scene.

This view is strongly supported by evidence on the relationship between perceived dress colours and the observers' implicit assumptions about the illumination of the scene on the photo (Chetverikov & Ivanchei, 2016; Hesslinger & Carbon, 2016; Toscani et al., 2016; Wallisch, 2017; Witzel et al., 2017). Moreover, seeing the real dress under normal viewing conditions (i.e. white light) does not yield any ambiguities: the dress is always seen as blue and black (Bach, 2015; Witzel, 2015), at least under neutral, broad-band illuminations (Hurlbert, Aston, & Pearce, 2016). An ambiguity in the colour perception of the real dress can only be achieved under particular illumination conditions (Hurlbert et al., 2016; Werner & Schmidt, 2016), which highlights the important role of the condition of illumination.

In order to account for individual differences in the perception of the dress, it has been speculated that these differences are due to individual differences in the subjective appearance of grey (Gegenfurtner et al., 2015; Lafer-Sousa et al., 2015; Winkler, Spillmann, Werner, & Webster, 2015). Individual differences in the subjective grey point are related to variations of colours along the daylight locus, which represents the colour changes of natural daylight (Bosten, Beer, & MacLeod, 2015; Chauhan et al., 2014; Witzel, Valkova, Hansen, & Gegenfurtner, 2011; Wuerger, Hurlbert, & Witzel, 2015). It has been proposed that these individual differences reflect different expectations, or *priors*, about the reference illumination, and that these different expectations could be related to the different interpretations of the dress colours (Gegenfurtner et al., 2015; Lafer-Sousa et al., 2015). One variant of this account suggested that the dress is related to different magnitudes of the *blue bias* (Winkler et al., 2015). According to the blue bias observers tend to judge a slightly bluish grey as completely grey (Weiss, Witzel, & Gegenfurtner, under review; Winkler et al., 2015; Wuerger, Hurlbert, & Witzel, 2015). However, existing evidence speaks against these ideas (Witzel et al., 2017; Wuerger et al., 2015).

According to still another view (Witzel, 2015; Witzel, Racey, & O'Regan, 2016; Witzel et al., 2017), the realism of the photo of the dress compels observers to spontaneously interpret the scene in one of two possible ways in order to make sense of the photo. The persistence of the perceived dress colours may be explained by observers getting locked into their initial interpretation and assumptions because they believe that this interpretation reflects the reality depicted on the photo. This idea is supported by observations according to which the perception of the dress may be shaped by prior experience with disambiguating images (Witzel, Racey, et al., 2016; Witzel et al., 2017) and one-shot learning (Drissi Daoudi, Doerig, Parkosadze, Kunchulia, & Herzog, 2017).

The observation that prior experience with disambiguating images influences the perception of the dress indicates the important role of top-down influences on the perception of the dress (Witzel, Racey, et al., 2016; Witzel et al., 2017). Further support for this idea comes from an fMRI study according to which white-gold seers have a stronger activation of brain regions critically involved in high-level processing, such as frontal and parietal brain areas (Schlaffke et al., 2015), from a study that found delayed visually evoked potentials in white-gold seers, which are indicative for the activation of higher cortical brain areas (Rabin et al., 2016), and from evidence for the role of beliefs about the real scene in the perception of the dress (Karlsson & Allwood, 2016).

One possibility is that the initial interpretation of the photo is as unpredictable as fluctuations in the perception of other bistable visual stimuli (Wexler, Duyck, & Mamassian, 2015). In this case, the individual variations of perceived dress colours would be random and not related to other phenomena. Alternatively, observers may differ more generally in the *perceptual strategies* by which they attribute features in two-dimensional images to the surfaces or to their illumination. In this case, the interpretation of illumination cues in the photo of the dress would be related to individual differences observed for other phenomena that involve the interpretation of cues to infer illumination and surface properties.

Individual differences have been observed for a whole range of such phenomena. First of all, substantial individual differences in colour constancy have been observed independently of the dress (Foster, 2011, for review; Granzier, Brenner, & Smeets, 2009; Granzier & Gegenfurtner, 2012, Fig. 13; Radonjic & Brainard, 2016; Witzel, van Alphen, Godau, & O'Regan, 2016). Moreover, a recent study found strong individual differences in gloss perception when stimuli were presented in two-dimensional photos, but not when observers saw the real three-dimensional stimuli (Hansmann-Roth, Pont, & Mamassian, 2015; Hansmann-Roth, Pont, & Mamassian, 2017). In the study of Lee and Smithson (2016) observers differed in whether they could use gloss information to discriminate changes in illumination from changes of surface colour. Häkkinen and Gröhn (2016) found pronounced individual differences in the way observers inferred shape from shading (Ramachandran, 1988). These individual differences could potentially be due to a fundamental difference in perceptual strategies concerning the interpretation of illumination and surface properties in two-dimensional images.

Here we tested whether the different kinds of individual variation discussed above are related to the differences in perception of the dress and to each other. For this purpose, we measured individual differences for different phenomena and tested whether they were correlated. In a first approach, we examined the relationship between individual variation in the perception of the dress and in colour constancy (see also Hurlbert et al., 2016). For this purpose, we reanalyzed two datasets collected in previous experiments,

one on the dress (Witzel et al., 2017) and the other on colour constancy (Witzel, van Alphen, et al., 2016).

In a second approach, we considered a wider range of phenomena more generally related to the interpretation of illumination and surface properties. For this we conducted an online study in order to obtain a large sample of observers. In contrast to experiments in the laboratory, experimental conditions, and in particular monitor calibration and colour rendering, cannot be fully controlled in online studies. Although this might be seen as a disadvantage, the discovery of the dress phenomenon happened online in the social media, which shows that at least the dress phenomenon does not depend on careful monitor calibration and experimental conditions. Moreover, subtle effects like the memory colour effect that were first measured under experimental conditions have actually been reproduced in several online surveys (Witzel, 2016). For these reasons, the online approach seemed promising to us.

The online experiment involved an extensive range of measurements on individual differences. These included individual differences in colour perception in ambiguous photos (dress, jacket), in the perception of gloss, illumination changes, and of shape from shading, in subjective grey points, and in colour naming. Preliminary results of the online study were presented at conferences (Witzel, Hansmann-Roth, & O'Regan, 2016; Witzel, Wuerger, & Hurlbert, 2016).

2. Experiments on dress and colour constancy

If the perception of the dress colours is a special case of colour constancy, individual differences in colour constancy might be directly related to the differences in the perception of the dress colours because they might be due to the same perceptual strategies. In two previous studies we had measured the perception of the dress (Witzel et al., 2017) and individual differences in colour constancy with largely the same sample of participants (Witzel, van Alphen, et al., 2016). This gives us the opportunity to revisit those two datasets and test the idea that individual differences in the perception of the dress are directly related to individual differences in colour constancy.

2.1. Method

2.1.1. Participants

Originally, 31 observers participated in the study on the dress (Witzel et al., 2017), and twenty observers in the experiment on colour constancy (Witzel, van Alphen, et al., 2016). Sixteen of those observers (13 women, 3 men; age: 24 ± 4 years) participated in both experiments. Red-green colour vision deficiencies were excluded through Ishihara plates (Ishihara, 2004). Apparatus was the same for all participants and in both studies. All details on the apparatus may be found in the previous publications (Witzel et al., 2017; Witzel, van Alphen, et al., 2016). This research was carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Informed consent about the experiments was obtained from each observer.

2.1.2. Experiment on the dress

In the experiment on the dress, participants were asked to adjust the colour they saw on the lace and the body of the dress. Observers were shown the dress on the left side of the screen, and a disk on the right side. The disk was presented in a random colour. Observers could adjust the colour of the disk along the green-red, blue-yellow, and lightness dimensions in CIELUV space. For this, six keys for the six poles of the three dimensions were available. The adjustment was implemented as a polar adjustment technique (Olkkonen, Hansen, & Gegenfurtner, 2008; Witzel et al., 2011).

2.1.3. Experiment on colour constancy

The measurements of colour constancy consisted in an asymmetric matching task with photo-realistic images. Four photo-realistic images of scenes were used as stimuli (see Fig. 4 in Witzel, van Alphen, et al., 2016). In each scene, there were twelve coloured objects (stones, tiles). One of the twelve objects was rendered in a test colour. The other eleven objects were rendered in random distractor colours. The remaining surround was achromatic and reflected the chromaticity of the illumination.

The colours in the scene were rendered based on the reflectances of Munsell chips (Kohonen, Parkkinen, & Jaaskelainen, 2006; Parkkinen, Hallikainen, & Jaaskelainen, 1989) and using one of two illuminants, a yellowish and a bluish illuminant with 5000 K and 12,000 K Correlated Colour Temperature, respectively (see Fig. 5 in Witzel, van Alphen, et al., 2016). The fifteen test colours consisted of three shades of grey, red, yellow, green, and blue, respectively.

Fig. 1 illustrates the stimulus display with an example. The stimulus display consisted of two images of the same scene and observers were asked to match the test colour across the two images. There were two conditions. In the colour constancy condition, observers were shown one of the four scenes under two different illuminations. One object was shown in the test colour in one and in a random colour in the other image. The random colour could be adjusted by the observer. The observer was asked to adjust the random colour so that it had the colour that corresponds to the one shown under the other illumination. The other condition was a control condition in which both images showed the scene under the same illumination. Apart from that, the adjustment task was the same as in the colour constancy condition and was used to assess observers' variability in colour perception that was not due to the illumination change. Participants could not adjust lightness, but only the chromatic dimensions red-green and blue-yellow using four keys.

Each test colour was adjusted under each type of illumination. There were thus two colour constancy conditions (yellow target illumination vs. blue comparison illumination, and blue target illumination vs. yellow comparison illumination), and two control conditions (adjustments under blue and under yellow illumination). Adjustments were repeated once across two sessions. Consequently, each observer provided 2 conditions (constancy vs. control) $*$ 2 target illuminations (blue vs. yellow) $*$ 15 test colours $*$ 2 repeated measurements, that is 120 adjustments.

2.2. Results

Fig. 2.a illustrates individual adjustments of the body and the lace of the dress. The perceived colours of the dress varied mainly along the first principal component in three-dimensional colour space (for details see Fig. 6 in Witzel et al., 2017). We therefore

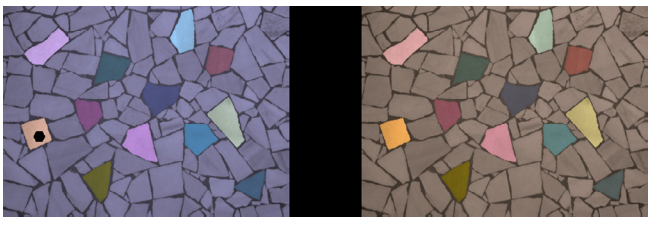

Fig. 1. Example of stimulus display in colour constancy experiment. The black dot indicates the test patch that observers matched to the corresponding orange-yellow patch in the other image. The illumination condition is yellow-to-blue (Y2B).

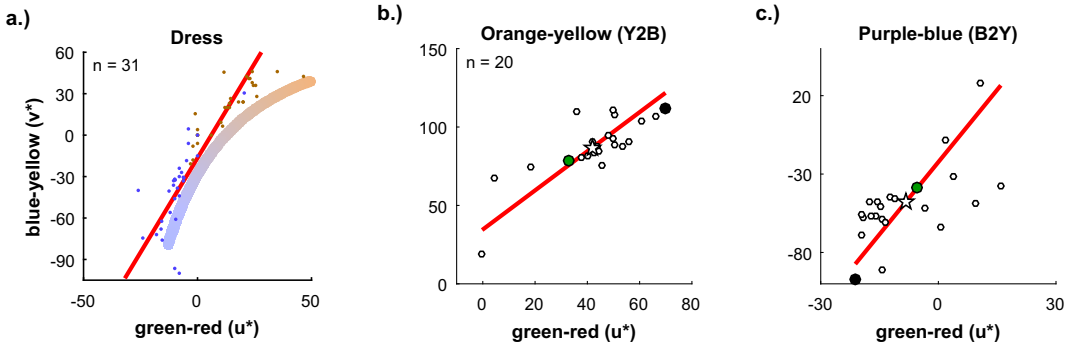

Fig. 2. Adjustments in the dress and in the colour constancy experiments. Adjustments are represented in CIELUV space, with the x- and y-axes referring to u^* and v^*. In all panels the red line corresponds to the first principal component that represents the main variation in the data. Panel a shows individual adjustments of the body (blue dots) and the lace (brown dots) of the dress for all 31 observers in Witzel et al. (2017). The coloured curve indicates the daylight locus. Panels b-c show asymmetric matches for two examples: orange-yellow when adjusting the colour under the blue illumination (panel b, Y2B = yellow to blue; cf. Fig. 1) and purple-blue when adjusting the colour under the yellow illumination (panel c, B2Y = blue to yellow). The black disk indicates the colour signal under the first illumination, the green disk indicates the colour under the second (bluish) illumination, assuming that the colours correspond to a typical natural surface (see text for details). White disks refer to average adjustments of each of the 20 participants in Witzel, van Alphen, et al. (2016), and the large white star indicates the grand average across participants.

focused on the scores along the principal component (*dress scores*). The lower the dress score the bluer was the body, and the higher the dress score, the yellower (more golden) was the lace. Fig. 2.b-c illustrates individual adjustments in the colour constancy experiment averaged across the two sessions for two example colours, namely orange-yellow adjusted under the blue illumination (as in Fig. 1) and purple-blue adjusted under the yellowish illumination in the constancy condition. The full original samples of participants (31 and 20, respectively) were used to determine general patterns through principal components. However, only the data for the 16 observers participating in both experiments could be compared between the two experiments.

2.2.1. Colour constancy performance

We first compared the dress scores to individual variation in the three measures of colour constancy performance proposed in the previous study (Witzel, van Alphen, et al., 2016). First, colour constancy was assessed in comparison to the prediction of the colour change based on natural surfaces. We calculated the distance between the colour signal predicted by a given surface reflectance and the adjustments provided by each observer, i.e. the distance between the white dots and the green dot in Fig. 2.b-c (*deviations from target predictions*). This measure is closely related to colour constancy indices (Arend, Reeves, Schirillo, & Goldstein, 1991) and like those it depends on knowledge about surface reflectances. Since the observer is most likely to know the properties of natural surfaces, we used prototypical reflectances of natural surfaces for target predictions (for the determination of prototypical natural surfaces see Witzel, van Alphen, et al., 2016). However, the observers do not necessarily have precise knowledge about target reflectances. For this reason, we also used two measures that assessed the observer's uncertainty about colours under illumination changes without assuming a target reflectance. Our second measure determined colour constancy as the precision with which an observer is able to identify the same colour over time. So, we calculated the distance between each observer's adjustment in the colour constancy condition of the first and second session, i.e. the variation of the white dots in Fig. 2.b-c over time (*intra-individual variation*). The third measure evaluated colour constancy in comparison to the overall expectation of the colour change across all observers. For this, we computed how similar each observer's adjustment was from the average of all observers, i.e. the distance between the white dots and the white star in Fig. 2.b-c (*inter-individual variation*).

For each participant, we averaged these colour constancy measures across all 15 test colours and both colour constancy conditions (yellow-to-blue and blue-to-yellow), and calculated correlations between these average colour constancy measures and the dress scores. Average deviations from target predictions and difference from the average (inter-individual variation) did not correlate with dress scores ($r(14) = -0.41$, $p = 0.11$; and $r(14) = -0.31$, $p = 0.24$), but differences across repeated measurements did ($r(14) = -0.66$, $p = 0.006$). The latter correlation implies that the more similar individual observers' adjustments were across the two sessions, the higher was their dress score. In other words, observers who were most reliable in their colour identification tended to see the dress as white-gold.

2.2.2. Systematic variation across individuals

Second, we examined systematic inter-individual variation of adjustments in the colour constancy experiment. In a first step, we isolated the variation that is common to all stimuli and both types of illumination changes in the colour constancy condition. For this purpose, we calculated the first principal component across the two dimensions of adjustments (green-red u^* and blue-yellow v^*) and across all 15 stimuli and 2 illumination conditions. The resulting first principal component of these overall 60 variables explained almost one third (32.5%) of the variance. The scores for this principal component were highly correlated with the dress scores and this correlation explained more than 58% of the total variance ($r(14) = 0.76$, $p = 0.0006$; Fig. 3.a). We verified that the correlation is statistically robust with the robust correlation toolbox (Pernet, Wilcox, & Rousselet, 2012). Pearson and Spearman correlations with boot-strapped confidence intervals, Bend correlations, skipped Pearson and Spearman correlations were all significant (all $p < 0.02$).

To better understand where this correlation comes from we then inspected whether this effect is related to particular target or illumination colours. For this, we calculated the principal component for the two adjustment dimensions (u^* and v^*) separately for each of the 15 test colours and each of the 2 illumination conditions. These principal components explained between 61% and 97% of the variance in two-dimensional space depending on the test colour and the illumination condition (cf. Figs. S1–S2 in Witzel, van Alphen, et al., 2016). We calculated correlations between the scores for each of these principal components and the dress score.

The variance explained by the single correlations is shown in Fig. 3.b. The correlations for the following conditions were signifi-

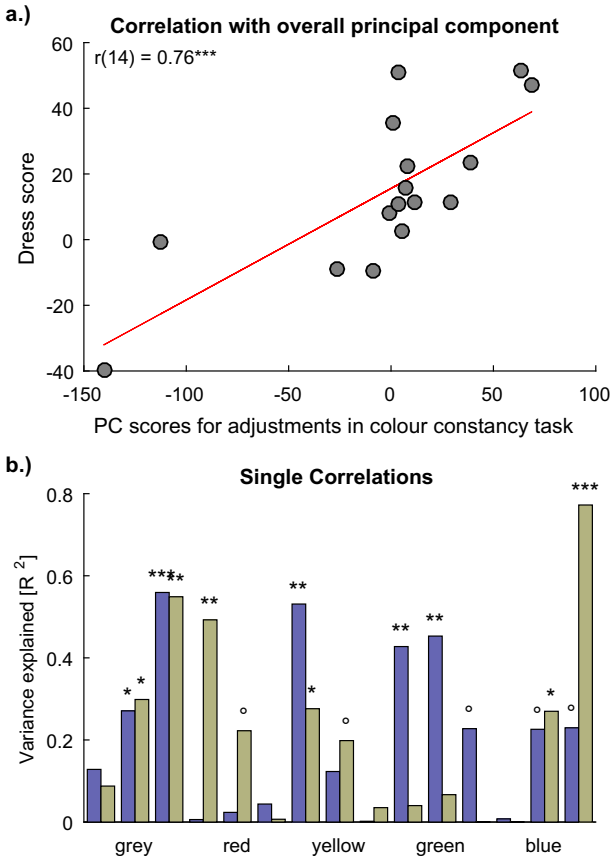

Fig. 3. Correlation between dress scores and individual differences in colour constancy. Panel a illustrates the correlation between dress scores (y-axis) and the scores of the first principal component of asymmetric matches across all observers and stimuli (x-axis). The regression line is shown in red. Panel b shows the explained variance for correlations between dress scores and principal component scores separately for each of the 15 stimuli and each of the two colour constancy conditions in the asymmetric matching task. Blue bars correspond to the yellow-to-blue and yellow bars to blue-to-yellow conditions. Symbols on top of the bars indicate significance of correlations. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\circ}p < 0.1$, ns = not significant.

cant after Bonferroni correction for 30 tests (alpha = 0.05/30 = 0.0017): For orange-yellow in the condition, in which the comparison illumination was yellow and the target illumination was blue (cf. Fig. 2.b; r(14) = 0.73, p = 0.0014); for purple-blue in the condition with the yellow target illumination (Fig. 2.c; r(14) = −0.88, p < 0.0001); for dark grey in both conditions (yellow-to-blue: r(14) = 0.75, p = 0.0009; blue-to-yellow: r(14) = −0.74, p = 0.001). In a multiple regression, these four components explained almost as much variance (56%, F(1,14) = 17.8, p = 0.0009) of the dress scores as the principal component of all measurements together (58%, see above).

Except for dark grey in the yellow-to-blue condition, the above three principal component scores varied from blue (negative score) to yellow (positive score). Hence, these principal components were aligned with the direction of the illumination change between blue and yellow (and vice versa). For dark grey in the yellow-to-blue condition the principal component was slightly rotated towards green and the scores varied from bluish purple (negative score) to yellowish green (positive score). The above correlations were positive for yellow-to-blue and negative for blue-to-yellow illumination changes. This implies that observers who tended to overshoot in the asymmetric matching (low scores in yellow-to-blue, high in blue-to-yellow), tended to see the dress as blue-black (negative dress score), and those who undershoot tended to see the dress as white-gold (high dress scores).

2.2.3. Shifts in the direction of the illumination change

Third, the correlation between dress scores and colour constancy adjustments may be due to differences in how observers attribute the colour change of a surface patch to the illumination or to the surface. If this is the case, the dress scores should be specifically related to individual differences in the direction of the illumination change. However, previously (Witzel, van Alphen, et al., 2016) we had observed that the principal components were not always oriented in the direction of the illumination change. To specifically examine variation along the illumination change, we projected the average adjustments of each observer onto the line that connects the white-points of the two illuminations. To capture the common variance across the 15 test colours and the 2 illumination conditions, we calculated the first principal component for these projected coordinates. This principal component explained 37.8% of the variance. The scores were significantly correlated with the dress scores (r(14) = 0.73, p = 0.001), and that correlation was statistically robust.

The colour of a given surface does not exactly follow the colour shift of the white-point. If observers take the colour change of a natural surface as their point of reference, a line connecting the colour signal under the two illuminations (i.e. connecting the black and green disks in Fig. 2.b-c) might be more relevant for individual variation than the change in illumination colour. For this reason, we also projected adjustments on the line that connects the colour signal of a natural surface under the two illuminations, and we calculated the first principal component of the projected coordinates across stimuli and illumination conditions. The principal component explained 40.4% of the variance. Its scores are also significantly correlated with the dress scores (r(14) = −0.63, p = 0.008, and the correlation was again statistically robust.

In sum, the projection of the adjustments on the direction of the illumination change (i.e. the white-point shift) and on the direction of the reflected colour signal excluded variation of the adjustments that were not aligned with these directions. Nevertheless, their principal components yielded correlations with the dress scores. In the case of the projections on the direction of the illumination change, the correlation (r(14) = 0.73) was close to the one obtained with the principal component of the original (i.e. non-projected) adjustments (r(14) = 0.76). This suggests that the variation of adjustments along the illumination change contains similar information about the dress scores as the original adjustments.

2.2.4. Control condition without illumination change

Finally, we also inspected the relationship between dress scores and the individual data in the control condition without change in illumination. We calculated the averages of the three measures of colour constancy for the control condition. None of them was correlated with the dress scores (all p > 0.73). Then, we calculated the overall principal component of adjustments across all 15 test colours and the two illuminations (yellow and blue) in the same way as for the colour constancy condition. That principal component explained 28.5% of the common variance. In contrast to the colour constancy condition, there was no correlation between the principal component scores for the control condition and the dress scores (r(14) = 0.04, p = 0.90). This result suggests that the observed relationship between dress scores and colour matching is specific for colour constancy conditions that involve an illumination change.

2.3. Discussion

The results provided us with three kinds of insight about the relationship between the perceived colours of the dress and individual differences in colour constancy.

2.3.1. Consistency of colour identification across illuminations

First, the analyses of the three measures of colour constancy performance indicated that observers who tend to see the dress as white-gold tended to be more consistent in colour constancy adjustments across repeated measurements. This result supports the idea that the dress is related to individual uncertainty in colour constancy. However, it is curious that there was no correlation between dress scores and inter-individual variation because inter-individual variation more directly reflects individual differences in colour constancy.

2.3.2. Biases in colour constancy

Moreover, we found evidence that observers who tended to undershoot their adjustments in the direction of the illumination change were inclined to see the dress as white-gold. This effect seemed to be specific to a few colours (dark grey, orange-yellow, purple-blue) under particular illumination changes (yellow-blue vs. blue-yellow). These colours were close to the daylight locus and to the direction of the illumination change in the experiment (blue-yellow). The projection of adjustments on the direction of the illumination change explained almost as much variance as the original (non-projected) adjustments.

These observations establish a first suggestion supporting earlier ideas that the dress is related to colour constancy (Brainard & Hurlbert, 2015; Gegenfurtner et al., 2015; Lafer-Sousa et al., 2015; Winkler et al., 2015; Witzel, 2015; Witzel et al., 2017). In particular, it has been speculated that the ambiguity of the dress may be due to the fact that the colours of the dress vary in the same hue direction as the colours of daylight illuminations and of the illumination of the scene in the photo (Gegenfurtner et al., 2015; Lafer-Sousa et al., 2015; Winkler et al., 2015; Witzel et al., 2017). Due to the overlap between surface and illumination colours, it is possible that some observers attribute colours in the photo to the fabric of the dress, while others attribute them to the illumination. Although we cannot pinpoint the precise origin of the observed correlations between perceived dress and colour constancy; they are generally in line with the idea that the individual differences in the perceived colours of the dress are related to how much observers attribute colours to the surface of the object or to its illumination.

2.3.3. General priors and the blue bias

Finally, two observations contradict the idea that the perceived colours of the dress are related to differences in the subjective grey point due to priors about the illumination (Gegenfurtner et al., 2015; Lafer-Sousa et al., 2015; Winkler et al., 2015). First, we did not observe any correlation between dress scores and the adjustments in the control condition without illumination change. This observation shows that the individual differences in the perception of the dress are not related to general uncertainties about colour (Gegenfurtner et al., 2015; Lafer-Sousa et al., 2015; Winkler et al., 2015), such as uncertainties about the subjective appearance of grey observed in previous studies (Bosten et al., 2015; Witzel et al., 2011). Instead, the different perceptions of the dress are specific to conditions that involve colour constancy.

Second, the observed correlations were not specific to the direction of colour change, but to the difference between under- and overshoot. If the differences were due to a difference in priors about the illumination (Gegenfurtner et al., 2015; Lafer-Sousa et al., 2015) or to a different magnitude of the blue bias (Winkler et al., 2015), there should be a consistent shift in one hue direction depending on the observers' prior or blue bias. However, our results suggested that blue-black seers overestimated and white-gold seers underestimated the effect of the illumination, no matter whether the illumination shifted from yellow to blue or from blue to yellow. Consequently, the individual differences in adjustments that are related to the dress go in either hue directions depending on the direction of the illumination change. This observation speaks against a fixed shift to either yellow or blue as predicted by the idea of different illumination priors or differences in the blue bias.

3. Online study on surface properties

The online study was mainly aimed at investigating general perceptual strategies about the attribution of two-dimensional cues to surfaces and illuminations. For this purpose, the online survey measured individual differences for a whole range of phenomena for which individual differences in the interpretation of surface properties were known, and tested whether they are correlated with the dress and with each other. We also measured individual differences in colour naming and in the subjective appearance of grey and examined whether they are related to the dress. The online study also allowed us to assess whether findings obtained under controlled conditions in the laboratory can be reproduced in online experiments despite the absence of display calibration and proper colour rendering. In a preliminary online survey we tested whether previous findings on the dress from our experiment in the laboratory (Witzel et al., 2017) can be replicated and further specified in an online study. Then, we conducted the main survey that compared the dress to other phenomena.

3.1. Method

3.1.1. Participants

For the preliminary experiment, 47 participants were recruited through emails and social media, and took part voluntarily. In the survey, observers were asked to indicate their age, gender, country of origin and whether they had colour vision deficiencies. 3 observers were excluded from the analyses because they indicated that they had colour vision deficiencies or that they didn't know whether they had colour vision deficiencies. The average age of the remaining 44 observers was 28.3 years (SD 6.4y) and 21 were women. Most observers in the final sample were from Serbia (45.5%), Germany (15.9%), and France 13.6%).

For the main experiment, 533 Observers were either recruited from an email list (CNRS., 2015) and participated voluntarily; or through a commercial online recruitment platform (Isis Software Incubator., 2015), and were paid for participation. 33 observers were excluded from the analyses because they reported that they had colour vision deficiencies or that they didn't know whether they had colour vision deficiencies. The average age of the remaining 500 observers was 31.9 years (SD 11.1y) and 239 were women. Most observers were from the UK (28.6%), France (18.6%), the USA (17.6%), India (5.6%), and Bosnia (4.2%).

3.1.2. Material

The experiment was implemented as a google form. A pdf print-version of the complete google form may be found in the Supplementary Material. The online study consisted of eight different kinds of stimuli and tasks.

3.1.3. Colour-ambiguous images

In the first kind of task, we measured the perceived colours for colour-ambiguous photos, such as the photo of the dress. If the differences in the descriptions of these images have the same origin they should be related one to another. The first column in Fig. 4 presents the colour-ambiguous images. Panel a shows the famous dress (Swiked, 2015). We will call this original photo "DressO" (dress original). The preliminary study only featured DressO. The main study also included the photo of a Jacket in panel d. This

photo appeared in the social media during the time following the dress and it was claimed that its colours are similarly ambiguous as the dress, leading some observers to describe it as blue and white and others as black and brown (poppunkblogger, 2016).

The main experiment also included two photos of another dress, shown in panels g and j. According to informal reports, the colours of the dress in panel g (henceforth *Dress2*) are ambiguously named as either green or blue (Witzel, 2012). However, this photo was taken under reddish tungsten light without white-point adjustment. The dress in the fourth image (henceforth *Dress3*) is the same photo as Dress2, but the reddish white-point has been corrected in Photoshop (Adobe Systems Incorporated., 2008). Each of the garments depicted in these images consists of two parts with different colours. The main part of the garments we shall call the "body", i.e. *dress body* and *jacket body*. In addition to the body, the three dresses have *lace*, and the Jacket a *logo*.

The previous experiment (Witzel et al., 2017) had shown that the relationship between perceived colours of DressO and the assumptions about the illumination can also be captured with colour naming data. This makes it very easy to measure the perceived colours of colour-ambiguous images in an online experiment. We presented the images one by one, and simply asked observers to describe the two parts of each piece of clothing using colour terms. For this they picked one of 14 colour terms. These were the eleven English basic colour terms pink, red, orange, yellow, brown, green, blue, and purple, and the three glossy colour terms bronze, silver, and gold. In the preliminary experiment there was no option for silver, but only the 13 other colour terms.

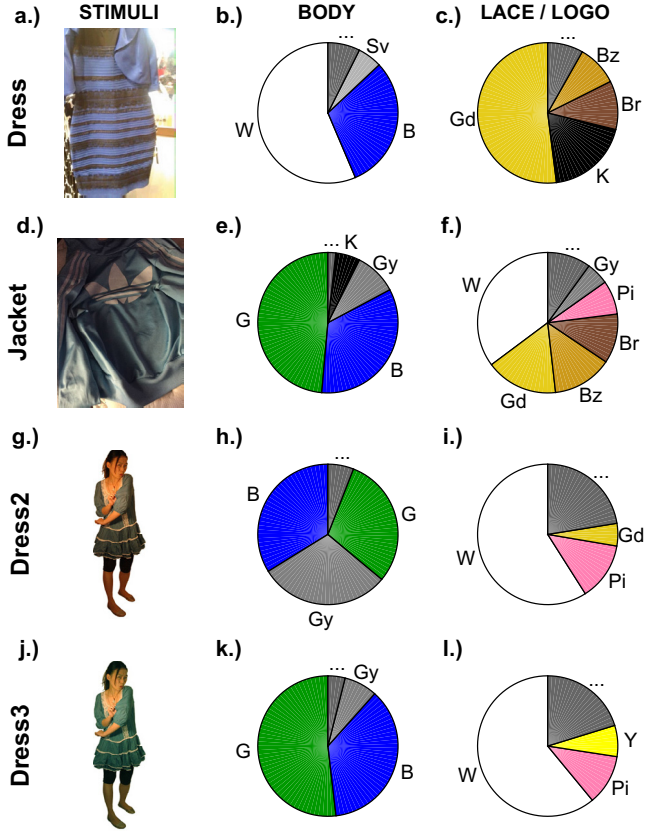

Fig. 4. Colour-ambiguous images in the main experiment. The four rows refer to DressO (a-c), the Jacket (d-f), Dress2 (g-i), and Dress3 (j-l), respectively. The first column (a,d,g,j) shows the colour-ambiguous images. The second and third columns represent the relative frequencies of colour terms used to describe the body (centre column) and the lace or logo (right hand column) of the dresses and jacket, respectively; W = White, Gy = Grey, K = Black, B = Blue, G = Green, Y = Yellow, Pi = Pink, Br = Brown, Gd = Gold, Sv = Silver, Bz = Bronze and '…'=other. See Fig. S1 for results from the preliminary experiment.

3.1.4. Assumed illumination

The second kind of task consisted of questions on the assumed illumination of DressO and the Jacket in order to extend and further specify previous findings (Witzel et al., 2017). For the original photo DressO, Witzel et al. (2017) had shown that estimation of the light that is illuminating the dress mainly varied between yellow and blue across observers. For this reason, the measurements of the colour of the assumed illumination were done along the yellow-blue dimension. Observers were asked to choose a number between 0 and 10, where 0 indicates yellow, 10 blue, and 5 colourless (white or grey). The assumed brightness of the illumination was determined by choosing a number between 0 for extremely dark and 10 for extremely bright. The respective scale was illustrated by a bar that continuously changed from yellow to blue (cf. Fig. 11.b), and from dark grey to white (cf. Fig. 11.a), respectively. Observers were asked to give these estimations for the light that was seen to be shining on the dress as well as for the light that was seen to be shining on the background of the dress.

We extended the set of questions used in the previous questionnaire-based experiment (Witzel et al., 2017). In one question we asked participants about the relationship between the dress and its background in the original photo (DressO). They could either answer "The dress is part of the background scene", "The dress is in the foreground, somehow separated from the scene in the background" or "Don't know". A second question asked whether the dress appeared to be illuminated by the same light as the background, and participants could either answer "Yes, the same light is shining on dress and background", "No, different lights are shining on dress and background", or "Don't know". A third question asked what kind of light was shining on the dress. For this question, descriptive statements were given, and participants could check the statements if they agreed. They could choose as many statements as they considered to be true. One statement said that the dress is in the shadow, another that the dress is illuminated by artificial light, a third that the dress is illuminated by the flash of the camera, and a fourth that the dress is illuminated by daylight/sunlight. Participants could also choose "Don't know" and "Other". The latter allowed them to add a free answer. The last question asked what light was shining on the background, and provided the answers "The background is illuminated by daylight/sunlight", "The background is illuminated by artificial light", "The photo is overexposed", "Don't know", and "Other". Just as for the third question, participants could choose as many answer statements as they agreed with.

Both the preliminary and the main survey included the questions about the colour and the source of the illumination for the photo of the dress. In addition, the main survey also included brightness and colour ratings and questions about the illumination of the Jacket. However, unlike the questions concerning DressO, those concerning the Jacket did not include questions about the light in the background of the Jacket. No questions about the illumination were asked for Dress2 and Dress3.

3.1.5. Gloss

In the third task, we examined individual differences in gloss perception. A recent study found strong individual differences in gloss perception when stimuli were presented in two-dimensional photos, but not when observers saw the real three-dimensional stimuli (Hansmann-Roth et al., 2015; Hansmann-Roth et al., 2017). These findings suggested that some observers interpret light spots as gloss that reflects the illumination, while others attribute them to lightness texture on a matte surface. In the online study, we measured the individual differences for those

stimuli to relate them to other phenomena. We also included an image that had previously been presented as a matte stimulus (Yang, Kanazawa, Yamaguchi, & Motoyoshi, 2015), but looked glossy according to spontaneous reports of several observers.

Fig. 5 illustrates the 25 stimuli used for the gloss ratings. First of all, this stimulus sample included eight images (4 photos and 4 rendered images) that yielded strong individual differences in previous studies of Hansmann-Roth and colleagues (2015, 2017). In those studies, objects had four lightness levels, black (K), dark grey (D), light grey (L) and white (W), and five levels of specularity that we will label 0 for matte (i.e. without specularity), 4 for highest specularity in this sample, and discrete values between 1 to 3 for the levels in between. All renderings were based on a Microfacet model using GGX to describe the microfacet distribution function and the Smith model to describe the shadowing and masking (Walter, Marschner, Li, & Torrance, 2007). Different objects were illuminated using different environment maps (Debevec, 1998). Photos were taken from real spray-painted surfaces with a calibrated camera (see Hansmann-Roth et al., 2017 for details on the production process).

The ambiguous stimuli used here were the light grey (L) and white (W) photos (p) and rendered images (r) with the two highest levels of specularity in the sample (3–4). The ambiguous stimuli are highlighted by being marked with a red 'a' in Fig. 5, and they are the photos pW4 (panel e), pL3 (f), pL4 (o), and pW3 (r) and the rendered images rL4 (h), rW4 (k), rL3 (m), rW3 (p).

In addition to the ambiguous images, we added seven control stimuli, which should produce particular ratings of glossiness. These stimuli were used to check whether observers understood the task and to assess how much ratings vary when stimuli are not ambiguous. Control stimuli are highlighted by a green 'c' in Fig. 5. The control stimuli included a "super specular" teapot (panel a: teapot5), a "super specular" potato (q: potato5), and a medium specular potato (j: potato3). Furthermore, there were the matte white (b: pW0), matte black (n: pK0), the medium specular dark grey (d: pD2), and the highly specular black (c: pK4) photos from Hansmann-Roth et al. (2017).

Finally, we included stimuli for which glossiness was not known a priori (highlighted by a yellow '?'). In particular, this set contained a matte (i: rG0) and a medium specular (g: rG2) rendered image with a light brownish/golden colour that coarsely corresponded to the apparent colour of the lace of DressO. There was also a metallic potato (metal potato) that had matte and specular aspects in that it had at the same time a sanded or brushed appearance and a global specularity due to the metallic material. There were the two stimuli from Yang et al. (2015)'s *Current Biology* article (s-t: "cb snail" and "cb teapot"). Finally, we included DressO and the Jacket. For DressO, we asked observers to rate glossiness of the body (u), the lace (v), and the bolero (w). The bolero is the jacket hanging over the shoulder of the dress, mostly visible in the upper right part of the photo. We included the bolero in the gloss judgements because preliminary to this study informal reports suggested that observers think the bolero is made of a different material than the dress itself and see it as glossier than the dress. For the Jacket, we assessed glossiness ratings of the body and the logo.

The fact that some observers see the lace of DressO as gold implies that they see at least the lace as glossy. It might be that the other observers, who see DressO as blue-black, attribute the gloss to the lighting conditions, such as a flash of the camera. For this reason, we asked observers to indicate what caused the gloss of DressO. The answer to this question was multiple choice. Observers could choose that "The light shining on the dress makes it appear shiny and glossy", that "The fabric of the dress is shiny and glossy", they could also choose both answers, or "Don't know". We asked the same question with respect to the Jacket.

3.1.6. Colour from gloss

Lee and Smithson (2016) found that only half of their observers were able to use the gloss information gathered from the two-dimensional images in order to decide whether colour changes were due to changes in illumination colour or to changes in surface colour. Since their task directly relates gloss and colour constancy to the distinction between surfaces and illumination, the task is a good candidate for investigating perceptual strategies used to accomplish this distinction.

For the fourth kind of task, we took the images from Fig. 4 in Lee and Smithson (2016) and created gif images that changed smoothly from one colour (e.g. red) to the other (e.g. blue) and back in an endless loop. Fig. 6.a-h provides a static description of the stimuli; the animated gif-images may be found in the Supplementary Material. Observers were shown the animated images,

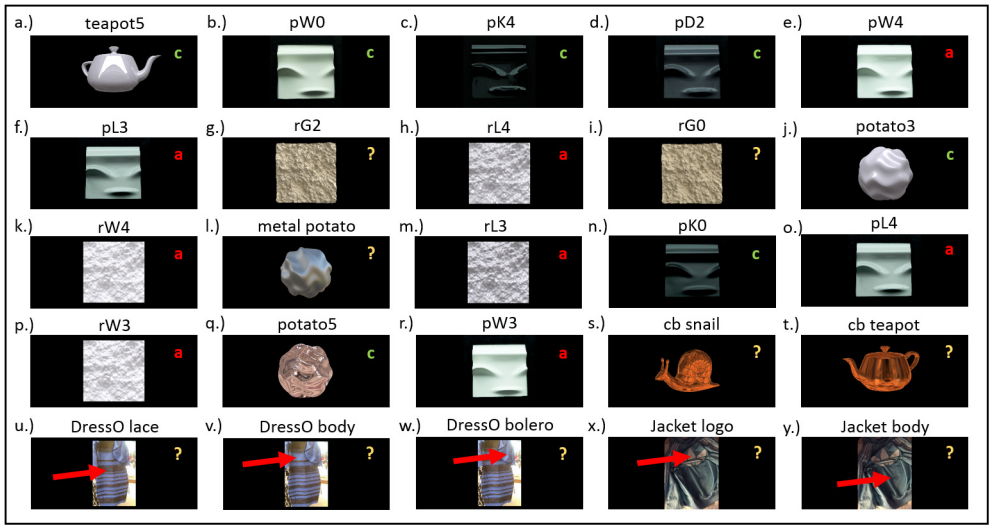

Fig. 5. Stimuli for gloss rating. The 25 stimulus images (a-y) are shown in the order of appearance in the online experiment. Stimuli that were assumed to be ambiguous based on previous studies (Hansmann-Roth et al., 2015; Hansmann-Roth et al., 2017) are highlighted by a red 'a'. Control stimuli are marked by a green 'c' and stimuli without any prior assumptions show a yellow '?'. The red arrows in panels u-y highlight different parts of DressO, i.e. the body (u), lace (v) and bolero (w), and of the Jacket, the body (x) and logo (y). In the experiment, ratings were done for the original images without any markers (i.e. without arrows, 'a', 'c' and '?'), which were only added for the purpose of illustration in this Figure. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and had to indicate whether the colour change was due to a change of the illumination or of the surface. Once you know how it works, the task is quite obvious: In case of an illumination change the specular highlight changes its colour because it reflects the illumination. In contrast, the specular highlight remains constant if the reflectance changes. The question is whether observers spontaneously use this information or not.

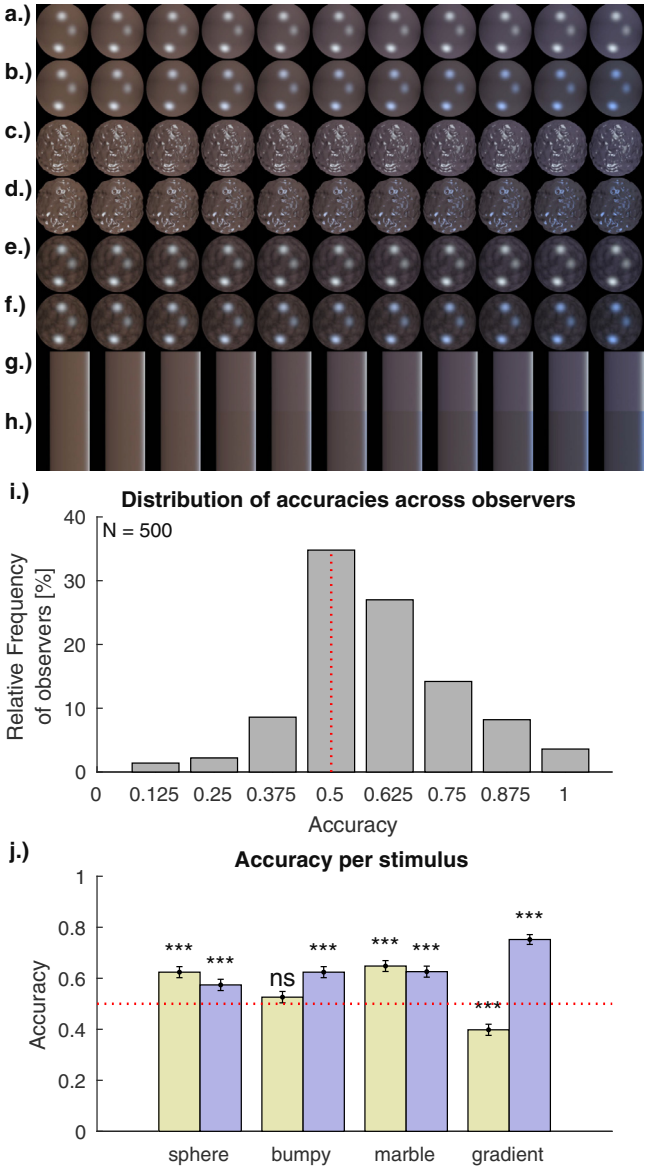

Fig. 6. Stimuli and results for colour-from-gloss. Panels a-h illustrate the eight types of stimuli. Each stimulus was an animated image that went through the eleven images shown from left to right and then back from right to left in an endless loop. The histogram in panel i illustrates the relative frequencies of observers (along the y-axis) for binned accuracy rates (along the x-axis). The accuracy rate for each observer was calculated across the eight stimuli. The red dotted line indicates the median. Panel j shows the average accuracy (along the y-axis) for each of the eight stimuli (along the x-axis). In this panel, accuracy rates have been calculated across observers. Yellow bars correspond to changes of reflectances (a,c,e,g), blue bars to changes of illumination (b,d,f,h). Here, the red dotted line refers to the chance probability of answering correctly (0.5). Error bars represent standard errors of mean. Symbols above the bars report significance of t-tests across observers comparing accuracy to chance level. $^{***}p < 0.001$, ns = not significant. Note that many observers were at chance level (panel i), and at the same time accuracy rates were significantly above chance level across observers for six out of eight stimuli (panel j). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1.7. Inverted waves

Häkkinen and Gröhn (2016) turned a photo of waves in the open sea upside-down (Fig. 7.a). While the upright photo is unambiguous, observers see completely different things in the inverted photo, such as water (rapids rather than waves), rocks, or a microscopic image of human tissue. The authors suggested that the individual differences in the perception of the inverted photo are due to the interpretation of two-dimensional cues about shape from shading (Ramachandran, 1988). It is also possible that some observers interpret light areas as highlights while others see them as part of the surface texture. We included this photo in the study because the differences in the interpretation of the two-dimensional cues to shape from shading and to specular highlights might be due to more general differences in perceptual strategies. As the fifth kind of task, we presented the inverted photo (Fig. 7.a) in our survey and asked participants what was shown on the photo. They could choose an option from water, organs, rocks, grass, tissue, or other. When choosing "other" they could enter a free answer.

3.1.8. Subjective grey

In a sixth kind of task, we added measurements of the subjective grey point. These measurements allow for testing the idea that the perception of the dress is related to differences in subjective grey point due to different priors and differences in the blue bias

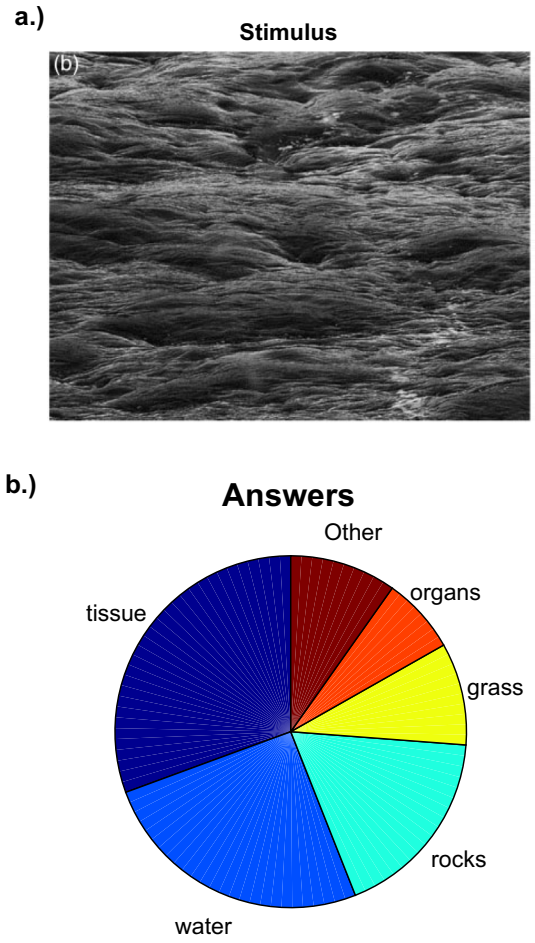

Fig. 7. Inverted Waves. Panel a shows the inverted waves image from Häkkinen and Gröhn (2016). The pie chart in panel b illustrates relative frequencies of responses. Note that relative frequencies were high for all response options indicating the lack of agreement across observers. Turn the image in panel a upside-down to see the original upright photo in which everybody sees waves without any ambiguity.

(Gegenfurtner et al., 2015; Lafer-Sousa et al., 2015; Winkler et al., 2015). To measure the subjective appearance of grey, we simply presented a grey and a bluish disk (Fig. 8.a), and asked observers which disk looked most grey. The grey and the slightly bluish disk were shown side by side, labelled as option A and B, respectively. Observers could choose either A or B as the "most grey" one.

3.1.9. Memory colour effect

We measured memory colour effects in a seventh kind of task. The memory colour effect corresponds to the influence of knowledge about an object's typical colour on the subjective appearance of grey (Hansen, Olkkonen, Walter, & Gegenfurtner, 2006; Olkkonen et al., 2008), and is related to individual variation in the subjective appearance of grey along the daylight locus (Bosten et al., 2015; Witzel et al., 2011). Individual differences in memory colour effects seem to be modulated by differences in the susceptibility of perception to cognitive effects (Witzel, O'Regan, & Rothen, 2016). The perceptual interpretation of the photo of the dress involves inferences beyond the sensory information of the image. For this reason, we wanted to test whether the perception of the dress can be explained by the different degrees to which observers let prior knowledge and assumptions influence their interpretation of the sensory information. To measure memory colour effects, we presented a grey and a bluish banana and asked which one looked most grey (Fig. 8.b). The banana looks grey when it is more bluish than colour-neutral objects, such as the disk (Hansen et al., 2006; Olkkonen et al., 2008; Witzel, 2016). We determined the amount of blue in the bluish disk (see Subjective grey) and banana according to the magnitude of the shift in perception due to the memory colour effect for the banana as measured by Olkkonen et al. (2008). The rational and features of the stimuli and the task used here have been described in more detail before (Witzel, 2016; Witzel, Olkkonen, & Gegenfurtner, 2016).

3.1.10. Colour naming

In an eighth kind of task, we examined the relationship between colour descriptions of the 4 ambiguous images (DressO, Jacket, Dress2, and Dress3) and individual differences in colour naming. A colour name such as "blue" refers to a colour category, e.g. the ensemble of all blue colour shades. Adjacent categories, such as blue and white have a boundary at which gradually changing colours, e.g. blue that increases in lightness, change category membership such as from blue to white. The precise location of these category boundaries varies across individuals (Olkkonen et al., 2010; Witzel & Gegenfurtner, 2013). If the colours on the colour-ambiguous photos are close to the category boundaries, this might be the reason why observers choose different colour terms to describe the colours on the photos. For example, some observers might see the same light bluish colour on the body of DressO, but one observer might call it blue while another calls it white. To test this idea, we assessed category boundaries that are relevant for the descriptions of the colour-ambiguous photos. Since we wanted to separate gloss perception from colour naming, we focused on the category boundaries that correspond to basic colour terms i.e. that exclude the metallic colour terms (gold, silver, and bronze).

In particular, we considered that the blue-white boundary (bw) would be relevant for the body and the brown-yellow boundary (bry) for the lace of DressO. For the body of the Jacket, of the Dress2, and of the Dress3, we measured a dark and a light boundary between green and blue (gb1 and gb2, respectively). Moreover, the scores for the Jacket, Dress2 and Dress3 had in common that one pole corresponded to a white logo or lace, respectively, which was contrasted to gold-brown or gold-bronze at the other pole of the score. For this reason, we also included a transition between brown and white.

In order to measure the category boundaries we rendered eleven colours that gradually transitioned from an unambiguous example of one category, e.g. blue, to an unambiguous example of the other, adjacent category, e.g. green. This was done for each of the five category boundaries. The resulting five sets of colour transitions are illustrated by Fig. 9. Observers were asked to pick the colour that is as much part of one as of the other category. The answers could range between 0, which was the colour closest to one category, e.g. the most greenish colour, to 10, which was the colour closest to the other category, e.g. the most bluish colour. These colour naming measurements were added later to the survey, so that only 405 of the 500 participants did the naming measurements.

3.1.11. Overall procedure

The experiment was divided into an introductory part and four main parts. The parts were separated by different "pages" of the survey. Observers were asked to finalise each part before starting the next one, and not to go back to previous pages.

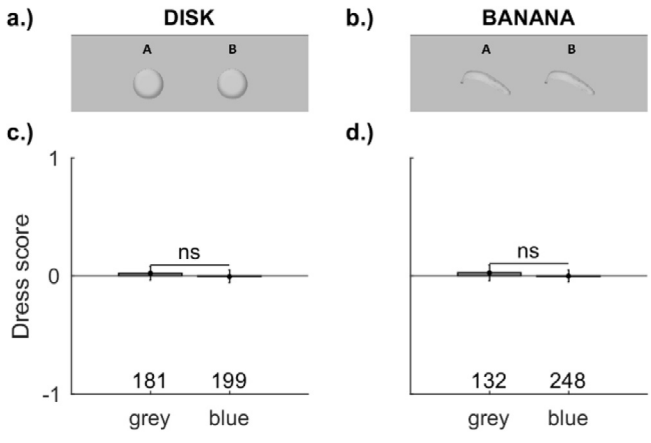

Fig. 8. Subjective grey points. Panels a and b illustrate the stimulus images used in the online experiment. Panels c and d show dress scores for DressO (along the y-axis) depending on which response option (grey or slightly bluish) observers chose (along the x-axis). Error bars correspond to standard errors of mean. The symbols on top of the bars indicate significance of an independent *t*-test (ns = not significant). The numbers at the bottom of the graphs report the frequencies of observers who chose that response option. A corresponding Figure for the preliminary experiment is provided by Fig. S7 in the Supplementary Material. Note that dress scores do not differ depending on subjective grey points.

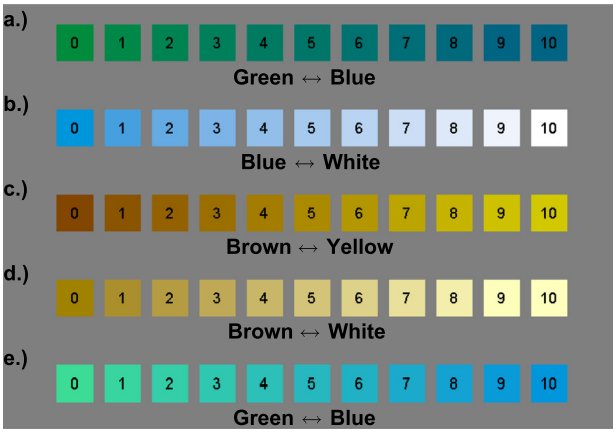

Fig. 9. Stimuli for colour naming. Panels a-e show transitions between different colour categories, namely green and blue (a,e), blue and white (b), brown and yellow (c), and brown and white (d). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The introductory part consisted of short instructions and the questions about participant characteristics (age, gender, etc.). At the end of this introductory part the question measuring the subjective grey point using the grey disk was presented.

The first main part was about the colours of DressO. It started with instructions that explained which part of DressO is called body and which part lace. These instructions did not involve any colour descriptions, but were done with arrows pointing to the respective areas of DressO. Observers indicated the colours first of the lace, and then of the body of DressO. At the end of this part, observers also responded to the inverted-waves task.

The second part was about the illumination of DressO. It started with instructions that explained and visualised what was considered to be its background. Then a section with the illumination estimations followed. First, observers judged the brightness of the illumination of the background and the dress. The section continued with the questions about the colour (yellow vs. blue) of the illumination of background and dress (in this order). A section with the five questions about the illumination followed after the illumination estimations. This part finished with the question measuring the memory colour effect using the grey banana.

The third part consisted of the gloss ratings. It started with an explanation of gloss, which showed example images that illustrate the difference between matte and gloss and that anchored the minimum (0) and maximum (10) of the scale with examples of completely matte and completely glossy objects. Then observers provided the 25 glossiness ratings. After the three ratings for DressO (lace, body, bolero), they also answered the question about the source of gloss (fabric or light), and after the two ratings for the Jacket (logo, body) they answered the question about gloss source for the Jacket.

The fourth part introduced the two images of the second dress (Dress2 and Dress3), and asked to describe their colours. It also included the five measurements of colour category boundaries. At the end of the survey, participants were provided with a link to a debriefing page and the contact details of the experimenter (the first author CW) for further questions.

3.2. Results

3.2.1. Colour-ambiguous images

The pie charts in Fig. 4 illustrate the frequencies of colour descriptions of the colour-ambiguous images and Table S1 provides the detailed numbers and proportions. Fig. 4.b-c and Fig. S1 provide the results for DressO in the main and in the preliminary experiments, respectively. In both experiments, slightly more than half of the observers saw the body of DressO as white, and most of the remaining observers saw it as blue. The lace was described as gold by more than half of the observers, and black, brown and bronze by most of the remaining observers. The body of the Jacket (Fig. 4.e) was mostly perceived as green or blue, the logo (Fig. 4.f) as white or in yellowish colours (gold, bronze, brown). The body of Dress2 (Fig. 4.h) yielded mostly responses of blue, grey, or green. The lace of Dress2 (Fig. 4.i) elicited mainly answers of white, pink, and several yellowish colour descriptions. The body of Dress3 (Fig. 4.k) was mainly seen as green, blue, or grey, and its lace (Fig. 4.l) as white, pink, or yellowish.

Our prior experiment (Witzel et al., 2017) had shown that a dress score based on colour naming bears the same relationship with estimations of illumination as do continuous measures of perceived colours obtained with colour adjustments. Hence, we translated the colour names describing the dress into a dress score to compare these descriptions to other phenomena.

For the preliminary experiment, the dress score for DressO was calculated based on the conceptual similarity between colour terms and the variation of perceived colours according to the colour adjustments measured previously (Witzel et al., 2017). This is illustrated by Fig. 10: The score is calculated as the sum of points for the body and the lace. A negative point is counted when observers answered "blue" for the body and "black" for the lace. A positive point was counted for answers "white" and "gold" for the body and lace, respectively. Other colour descriptions were given values between -1 and 1 depending on how light and how bluish-yellowish they are. For example, purple for the body and brown for the lace were scored -0.5, because they were dark and purple was close to blue. Hence, the negative pole of this score is -2 for blue and black, and the positive pole 2 for white and gold. Other combinations give values in between. For example, blue-gold would be $-1 + 1 = 0$. The dress score was significantly positive ($M = 0.52$, $t(43) = 2.1$, $p = 0.04$) because most observers saw DressO as white-gold.

Since the main experiment involved a large amount of data, it is possible to develop a more objective way to determine the dress scores. For this, we coded the answer to each of the 14 colour names binarily, i.e. one if observers chose it and zero if not. We did this for the lace and the body separately, and then concatenated the two datasets, resulting in 30 binary variables with 500 cases corresponding to the participants. To capture the covariation between those binary variables, we calculated the first principal component, which represents most of the common variance across all possible colour combinations. We used the scores of the first principal component as the dress or jacket score for the four colour-ambiguous images in Fig. 4.a,d,g,j. For the data of the main experiment, the dress scores based on the binary principal components were almost perfectly correlated to the conceptual dress scores ($r(498) = 0.99$, $p < 0.000001$). This shows that the approach based on principal components provided almost the same results as the conceptual coding, and it retrospectively validates the conceptual coding.

Table S2 in the Supplementary Material illustrates the resulting binary principal components in detail. The binary principal component for DressO (Fig. 4.a-c) contrasts white-gold to blue-black and explains 47.3% of the variance of the 30 variables. For the Jacket (Fig. 4.d-f), the principal component contrasts green-gold to blue-white and explains 35.0% of the variance. The principal component for Dress2 (Fig. 4.g-i) explains 32.5% and contrasts grey-pink to blue-white. Finally, the principal component for Dress3 (Fig. 4.j-l) explains 35.1% of the variance and contrasts green-pink to blue-white.

Table S3 in the Supplementary Material summarises the correlations between the principal component scores of the four colour-ambiguous images. The scores of the Jacket were highly signifi-

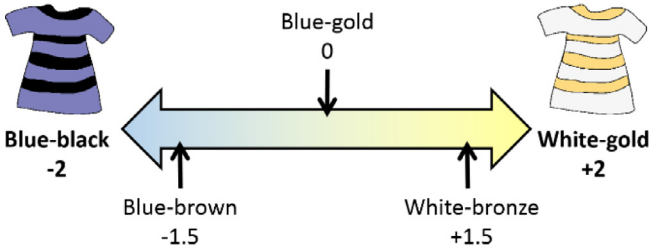

Fig. 10. Illustration of dress scores. This diagram illustrates how dress scores were calculated, with the blue-black combination forming the negative pole (−2) on the left, and white-gold the positive pole of the score (+2) on the right side. A score of zero meant that the two colour terms were completely ambiguous along the blue-black vs. white-gold dimension. This is the case for example for blue-gold. Details on principal component scores used in the main experiment as dress and jacket scores are presented in Table S2. Note that the principal component for DressO was very similar to the conceptually derived dimension of the dress score illustrated here. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cantly correlated with those for Dress 2 ($r(498) = 0.34$, $p < 10^{-14}$) and Dress 3 ($r(498) = 0.33$, $p < 10^{-13}$), and the correlation between the scores for Dress2 and Dress3 was still stronger ($r(498) = 0.54$, $p < 10^{-38}$). These correlation show that the colour descriptions of the Jacket, Dress2 and Dress3 are systematically related and hence likely to have similar determinants. In contrast, the scores of DressO were not correlated to any of the scores of the three other images (max. $r(498) = -0.02$, min. $p = 0.62$). This suggests that the individual differences in the colour descriptions of DressO have nothing in common with the three other photos.

3.2.2. Illumination estimation

In order to test for the relationship between the perceived colours of DressO and the estimated illumination, we calculated correlations between dress scores (Fig. 10.c) and illumination ratings for brightness (Fig. 11.a) and colour (Fig. 11.b), respectively.

Fig. 11.c–h illustrate the results from the main experiment; Fig. S2 in the Supplementary material provides the corresponding diagrams for the preliminary experiment. Note that in those diagrams we dichotomised the dress score for illustration purposes (because a scatter plot is not very helpful due to the discrete answer options). Observers with a score < 0 were assigned to the blue-black (BK, blue bars) and observers with a score > 0 to the white-gold (WG, yellowish bars) group. Significance rates in the plots refer to t-tests comparing the two groups. They reflect the main results found with correlations. All correlations are summarized in Table S4 of the Supplementary Material.

The previous experiment (Witzel et al., 2017) revealed a correlation along the yellow-blue dimension according to which the dress is perceived blue-black (white-gold, respectively) when observers assume that the light shining on the dress is more yellowish (more bluish, respectively). This correlation between dress score and the assumed colour of the illumination of the dress

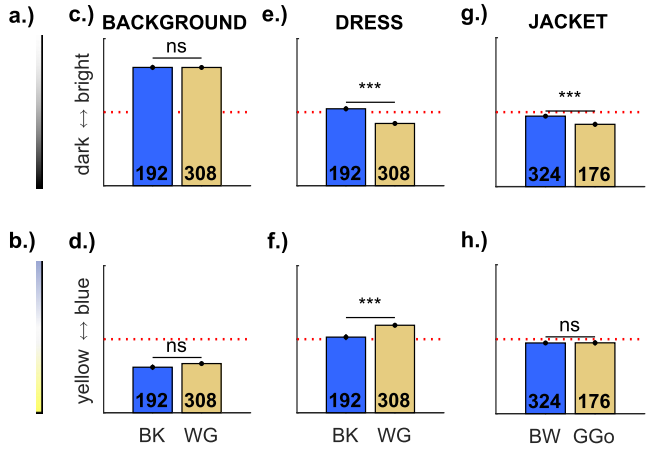

Fig. 11. Estimation of illumination of DressO and the Jacket. The first column (a-b) reproduces the bar-type visualizations that we gave observers to indicate their choice along the brightness and yellow-blue gradations. The second and third column show the estimations for the illumination in the background of the dress (c-d) and for the light taken to be shining on the dress (e-f). The last column illustrates the estimations of the light taken to be shining on the Jacket. Blue bars refer to groups with scores below and yellowish bars to groups with scores above zero. BK = Blue-black (negative score for dress), WG = White-gold (positive score for dress), BW = Blue-white (negative score for Jacket), GGo = Green-Gold (positive score for Jacket). Error bars correspond to standard errors of mean. Symbols above the bars report significance of t-tests across observers comparing observers with dress and jacket scores above and below zero. $^{***}p < 0.001$, ns = not significant. For results in the preliminary experiment see Fig. S2. Note that the correlations between perceived dress colours and estimations of the brightness and colour of the illumination of the dress (e-f) replicate the laboratory findings (Witzel et al., 2017). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

was replicated in the main online experiment (Fig. 11.f; $r(498) = 0.21$, $p < 0.00001$), but it was not significant in the preliminary study (Fig. S2.f, $r(42) = 0.18$, $p = 0.24$). Moreover, in both online surveys there was a significant negative correlation between dress scores and the estimated brightness of the light perceived to be shining on the dress (Fig. S2.e, $r(42) = -0.50$, $p = 0.0005$ and Fig. 11.e, $r(498) = -0.27$, $p < 10^{-9}$). This correlation implies that when observers take the illumination of the dress to be lighter, they become more likely to see the dress as blue and black rather than white and gold. Like those for the dress, the brightness estimations for the illumination of the Jacket were negatively correlated with the jacket scores ($r(498) = -0.17$, $p = 0.0002$), but no such correlation was found for the yellow-blue estimations of the illumination of the Jacket ($r(498) = 0.006$, $p = 0.89$). Observers who estimated the illumination to be bright in the photo of the Jacket, were more likely to see it as blue-white. Note that the observed correlations are still significant after correcting for multiple testing (4 estimations) according to Bonferroni (alpha = 0.0125).

We combined the brightness and colour estimations from the main experiment in a multiple regression to predict dress and jacket scores, respectively. The regression for the dress explained 9.2% of the variance and was highly significant ($F(2497) = 25.0$, $p = 10^{-11}$). The one for the Jacket explained only 2.7% of the variance and was also highly significant ($F(2497) = 7.0$, $p = 0.001$).

Finally, there was no correlation between the dress scores and any of the two estimations of the illumination in the background (all $p > 0.41$; see Table S4). Instead, the estimation of the brightness (Fig. 11.c and Fig. S2.c) and colour (Fig. 11.d and Fig. S2.d) of the illumination in the background were very similar for observers with negative and positive dress scores. For the brightness judgements, the histogram of the ratings (Fig. S3.a) shows that participants unanimously judged the background colour as bright, which implies that there might not be any systematic individual differences in the judgement of brightness for the background illumination.

3.2.3. Questions about illumination

For question 1 about the foreground-background relationship, and for question 2 about the sameness of the illumination of dress and background we discarded "Don't know" answers from the analyses. We compared dress scores between observers who answered yes and those who answered no in a two-tailed *t*-test for independent samples (see Table S5 for details). Fig. 12 illustrates the most important relationships between the answers to the questions and the dress score for DressO (first row: a-e) and jacket score (lower row: f-i) in the main experiment. Corresponding diagrams for the results from the preliminary experiment are provided in Fig. S4 of the Supplementary Material.

The most consistent effect appeared for the question of whether DressO was in the shadow or not. Observers who answered "yes" to this question had significantly higher dress scores than those who did not. This was the case in the preliminary (Fig. S4.a, $t(42) = 2.1$, $p = 0.04$) and in the main experiment (Fig. 12.a, $t(498) = 8.0$, $p < 10^{-14}$). These results replicate those found in the previous questionnaire-based experiment (Witzel et al., 2017). They show that observers who believe the dress is in the shadow tend to see it as white-gold rather than blue-back.

Further evidence for this idea is provided by answers with the complimentary pattern: observers who believe the dress is under a strong light tend to see the dress as blue-black rather than white-gold. In particular, observers who assumed the dress was illuminated by the flash of the camera (Fig. 12.b) or by daylight (Fig. 12.c) yielded significantly lower dress scores than those who did not ($t(498) = -4.2$, $p < 0.0001$ and $t(498) = -3.0$, $p = 0.003$). These results did not appear in the preliminary experi-

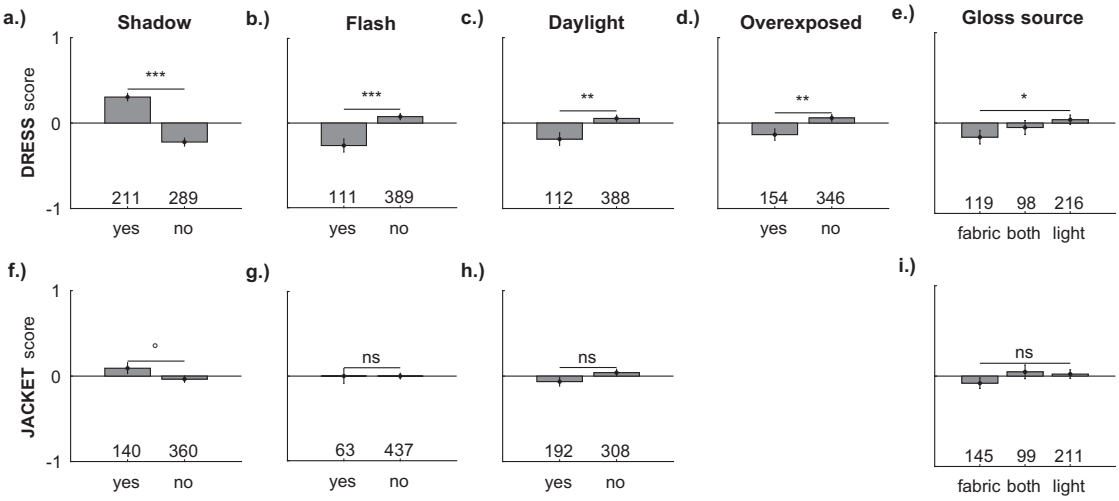

Fig. 12. Questions about the illumination. The first row (a-e) shows the dress scores for DressO (along the y-axis) as a function of the observers' answer to the respective questions concerning the illumination. The second row (f-i) illustrates the same kind of results for the jacket scores. The first column ("shadow") refers to the question of whether the dress (a) and the Jacket (f) were in the shadow ("yes"=in the shadow). The second column ("flash") distinguishes between observers that believed the dress (b) and the Jacket (g) were in the flash of the camera. In the third column ("daylight") the question was whether observers thought the dress (c) and the Jacket (h) were in daylight. Panel d ("overexposed") concerns the question of whether the image of DressO was overexposed (there was no such question for the Jacket). The last column refers to the question about the source of gloss (reported in the section on gloss). In this question observers could indicate that they believed the gloss was due to the fabric, the light or both. Numbers at the bottom of the diagram indicate the number of observers that gave each answer option. Error bars correspond to standard errors of mean. Symbols above the bars report significance of t-tests comparing dress and jacket scores between observers who answered yes and No. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\circ}p < 0.1$, ns = not significant. Note the significant results for DressO (first row), but not for the Jacket (last row).

ment (Fig. S4.b-c), but they were still significant after a Bonferroni correction for the 8 questions in the main experiment (alpha = 0.006). Moreover, the finding on the question about daylight is in line with results in the previous study (Witzel et al., 2017).

For DressO, we also asked three questions about the illumination in the background of the photo and about overexposure. The data from the main experiment showed that observers who thought the photo was overexposed tended to have a significantly more negative (towards blue-black) dress score than those who did not think the photo was overexposed (Fig. 12.d, $t(498) = -2.6$, $p = 0.009$). This result is also significant after Bonferroni correction for the 3 questions (alpha = 0.017), and it is in line with the general idea that the dress is perceived relative to the overall illumination. However, there were no other significant results concerning the light in the background in the preliminary (all $p > 0.55$) and the main experiment (ps > 0.45).

Finally, in contrast to DressO, there was no evidence for a relationship between the perception of the Jacket and the questions about the illumination of the Jacket (Fig. 12.f-i). The difference between jacket scores for those who thought the Jacket was in the shadow and those who did not was close to significance (Fig. 12.a, $t(498) = 1.8$, $p = 0.07$); but not when significance was Bonferroni corrected. However, it might be interesting to keep this result in mind given the observation of strong effects of assumptions about the shadow on the perception of DressO. This observation is also in line with the strong effect of brightness estimations on the perception of the Jacket (Fig. 11.f).

3.2.4. Gloss

Fig. S5 in the supplementary material provides histograms of gloss ratings for each of the 25 stimuli in Fig. 5. Fig. 13.a allows for comparing central tendencies of gloss ratings across stimuli.

Several stimuli yielded very consistent responses across observers. For example, the photo of the matte white surface (pW0, cf. Fig. 5.b and Fig. S5.b) and the golden matte rendered surface (rG0, Fig. 5.i and Fig. S5.i) were unambiguously perceived as matte (first two bars in Fig. 13.a), while the super-specular teapot (teapot5, Fig. 5.a and Fig. S5.a) and in particular the super-specular potato (potato5, Fig. 5.q and Fig. S5.q) were clearly perceived as glossy (last two bars in Fig. 13.a) by almost all observers. The consistency of ratings across observers is shown by the fact that the distributions of the gloss ratings for these stimuli show clear peaks either at the lowest (Fig. S5.b,i) or at highest end of the scale (Fig. S5.a,q), respectively.

Fig. 13.b-c illustrate the gloss ratings for the snail and the teapot (cb snail and cb teapot in Fig. 5.s-t) from the infant study (Yang et al., 2015). These stimuli yielded systematic individual differences in that their gloss ratings were positively correlated across participants ($r(498) = 0.65$, $p < 10^{-67}$). Nevertheless, these stimuli yielded comparatively narrow distributions of gloss ratings and were rated amongst the glossiest objects (cb snail and cb tea in Fig. 13.a), directly after the super-glossy teapot (Fig. 5.a) and potato (Fig. 5.q). This strongly contradicts the assumption in the infant study that these objects are completely matte.

In contrast to those unambiguous stimuli, there were also ambiguous stimuli for which observers disagreed in their gloss ratings. For example, this was the case for the photo of a glossy white surface (pW4, cf. Fig. 5.e) and a rendered white surface (rW4, cf. Fig. 5.k), for which Hansmann-Roth and colleagues (Hansmann-Roth et al., 2015; Hansmann-Roth et al., 2017) found strong individual differences. Although pW4 and rW4 were a photo of an object and a rendered object with highest specularity (in the sample), their mode ratings were only 2 and 6 (pW4 and rW4 in Fig. 13.a). More importantly, the gloss ratings for those stimuli featured broad distributions across the whole scale (cf. Fig. S5.e,k), which shows that observers rate these stimuli very differently.

In the Supplementary material, we discuss individual differences for other examples of control and ambiguous stimuli. In sum, the ambiguous gloss ratings show that there is strong individual variability for some stimuli. The observation that there are control stimuli that yielded ratings with a narrow distribution indicates that individual differences are not due to misunderstandings of the task, but are a feature of particular stimuli.

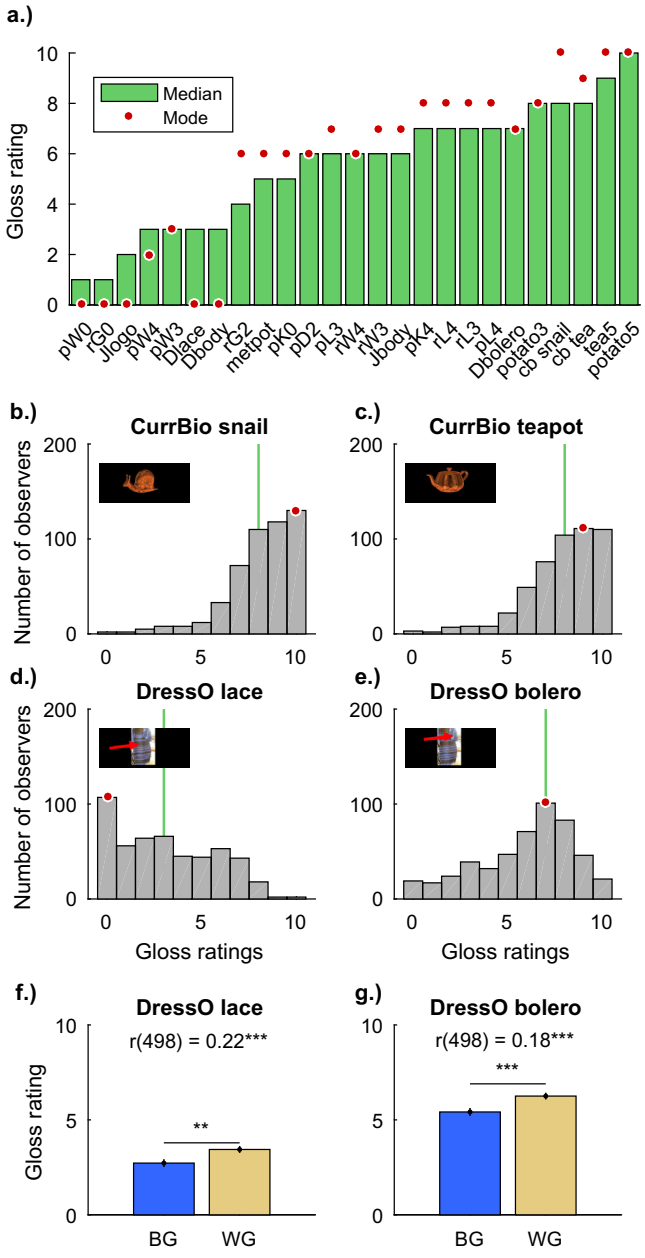

Fig. 13. Gloss ratings. Panel a compares central tendencies of gloss ratings across all stimuli. The y-axis represents gloss ratings. The green bars correspond to median, the red dots to mode gloss ratings calculated across the 500 observers. Stimuli are listed along the x-axis in ascending order of median ratings. Panels b-e provide histograms of gloss ratings for the snail and the teapot from the infant study (Yang et al., 2015) and for the lace and the bolero of DressO, respectively. These panels show the same data as Fig. S5.s-t, u, w). The x-axes of these four panels represent the scales (from 0 to 10) along which observers rated gloss. Values along y-axes give the number of observers that chose a particular gloss level for a given stimulus. Median and mode are shown as a green line and a red dot respectively. The last row (f-g) illustrates the relationship between gloss ratings and perceived colours of DressO. This is shown for the gloss ratings of the lace (f) and the bolero (g) of DressO. For illustration purposes, dress scores were dichotomized, resulting in groups with dress scores below (BK, blue bars) and above zero (WG, yellow). Gloss ratings are shown along the y-axis. Error bars correspond to standard errors of mean. Symbols above the bars indicate significance of t-tests comparing the two groups. The corresponding correlations reported in the main text are given at the top of the panels. $^{***}p < 0.001$, $^{**}p < 0.01$. Note that cb snail and cb teapot (b-c) are judged the glossiest after the two control stimuli tea5 and potato5 (a), and that the gloss ratings for the lace (d) and the bolero (e) were related to the perceived colour of DressO. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The colour-ambiguous images DressO and the Jacket also showed some ambiguities in the gloss ratings. The gloss ratings for the lace of DressO (Fig. 5.u) are illustrated by Fig. 13.d. The gloss ratings show a peak at matte (0) but an otherwise rather broad distribution between 1 and 8. The distribution of the gloss ratings for the body of DressO (Fig. 5.v and Fig. S5.v) was very similar to the one for the lace. In contrast, the bolero of DressO (Fig. 5.w) yielded much higher gloss ratings with a mode at 7 and a distribution spreading between 0 and 10 (Fig. 13.g). The gloss ratings for the three parts of DressO were significantly correlated (min. $r(498) = 0.29$, max. $p < 10^{-10}$).

The logo of the Jacket (Fig. 5.x) was rated as one of the most matte objects with a mode of 0 (Jlogo in Fig. 13.a and for details Fig. S5.x), directly after the matte white photo (pW0) and the matte golden rendering (rG0). In contrast the body of the Jacket (Fig. 5.y) had a mode at 7, but with a broad distribution (Fig. S5.y), indicating individual differences. The correlation between the ratings for the logo and the body of the Jacket was comparatively low ($r(497) = 0.11$, $p = 0.02$). At the same time, the logo yielded a positive correlation with the body and lace of DressO (min. $r(497) = 0.24$, $p < 10^{-7}$), and the ratings for the body of the Jacket were correlated to the body ($r(497) = 0.15$, $p = 0.001$) and the bolero of DressO ($r(497) = 0.36$, $p < 10^{-15}$). This indicates that there are some systematic individual differences in gloss ratings for DressO and the Jacket; but they are not straight forward to interpret.

In order to relate the individual differences in gloss ratings to DressO and the other colour-ambiguous images, we first calculated a *gloss score* that captures the common variance across all gloss ratings. The gloss score was simply the first principal component of the gloss ratings for the 25 objects, which explained 22.4% of the common variance (i.e. more than 5 times the variance of a single variable). However, this gloss score was not correlated to the dress and jacket scores for any of the four colour-ambiguous images ($r(497) = [-0.06, 0.07]$, all $p > 0.13$).

We then correlated gloss ratings for each single object with the dress and jacket scores that reflect the perceived colours. We focus on correlations that are significant after Bonferroni correction for 25 tests (alpha = 0.002). According to this criterion, the gloss ratings for the dress lace and the dress bolero were significantly positively correlated with the dress score of DressO ($r(498) = 0.22$, $p < 10^{-5}$ and $r(498) = 0.18$, $p = 0.00006$). This is illustrated by Fig. 13.f-g. According to these correlations, observers who saw more gloss in the lace and the bolero of the DressO also tended to see it as white-gold. This is in line with the idea that gold implies glossiness. At the same time, it is unexpected that dress scores were correlated to the gloss ratings for the dress bolero, but not to those for the dress body. There were also a few other unexpected correlations that we discuss in section "Gloss and colour-ambiguous images" of the Supplementary Material.

Fig. 12.e illustrates the relationship between the question about the source of gloss and the perceived colour of DressO. We discarded "Don't know" answers from the analyses of this question. We coded answers that attributed the gloss to fabric as -1, and those that attributed the gloss to the light of the dress as + 1. Answers that attributed gloss to both light and fabric were coded as 0. We calculated correlations across observers between these values and the dress scores for DressO. Dress scores were negatively correlated with assumptions about the source of gloss in the main experiment ($r(431) = -0.10$ $p = 0.03$) and in the preliminary experiment (Fig. S4.e, $r(34) = -0.35$, $p = 0.04$). A negative correlation implies that observers who tended to attribute the gloss of DressO to the fabric (-1) rather than the light (+1) were inclined to see the dress as white-gold (high dress scores). In contrast to

DressO, there was no such correlation between perceived colours and the assumed source of gloss for the Jacket ($r(453) = -0.06$, $p = 0.21$; cf. Fig. 12.i).

3.2.5. Colour from gloss

Fig. 6. i provides a histogram of overall accuracies in the colour constancy task. These accuracies were the average across the eight stimuli, calculated for each participant. The number of participants in a certain accuracy range is shown along the y-axis of Fig. 6.i. Average accuracy across the 500 observers was 59.7%, which is significantly above chance, i.e. above 0.5 ($t(499) = 12.7$, $p = 10^{-31}$). Nevertheless, the answers of 235 (47.0%) observers were equal or below chance level (≤ 0.5) across the eight stimuli, indicating that almost half of the observers could not do the task.

Fig. 6.j shows accuracies for each of the eight stimuli separately (see Table S6f or detailed results from comparing proportions correct with chance). Accuracy for the bumpy sphere that changed in reflectance (Fig. 6.c and third bar in Fig. 6.j) was not significantly above chance ($M = 52.6\%$, $t(499) = 1.2$, $p = 0.25$), which indicates that participants could not identify the correct answer (illumination change) for this stimulus. The gradient that changed in reflectance (Fig. 6.g and seventh bar in Fig. 6.j) was even significantly below chance level (39.8%, $t(499) = -4.7$, $p = 0.00001$), indicating that participants systematically picked the wrong answer (i.e. illumination change) for this stimulus. However, all other stimuli yielded accuracies significantly above chance level ($M = 57.4\%$-75.2%, $t(499) = 3.3$–13.0, all $p < 0.001$). In order to relate the performance in this task to the colour-ambiguous images, we correlated average accuracies per observer with the respective scores (see Table S6 for detailed results). The average accuracies were not correlated with DressO, Jacket, Dress2, and Dress3 (all $p > 0.13$).

Average accuracies might not well reflect the pattern that distinguishes individual differences in perceptual strategies. Some observers might tend to attribute colour changes to the illumination, while others might be more inclined to attribute colour changes to the surfaces. These different tendencies would not be reflected in accuracies because for some stimuli it is correct to attribute the colour change to the illumination, and for others it is incorrect. In the Supplementary Material we provide further analyses of such tendencies and their relationship to colour-ambiguous images. However, none of them were significant. In sum, there is no evidence for a systematic relationship between this task and the perception of colour ambiguous images.

3.2.6. Inverted waves

The pie chart in Fig. 7.b illustrates the relative frequencies of answers to the inverted-waves image in Fig. 7.a. Most observers chose tissue (30.6%), water (25.4) or rocks (17.8%). 49 observers (9.8%) chose 'other'. Most observers of Häkkinen and Gröhn (2016) chose water or rocks. This difference with our study might be due to the fact that the answer "tissue" in our study may refer to both microscopic human tissue as in Häkkinen and Gröhn (2016) or other kinds of tissue, such as paper or fabric.

Fig. S6 in the Supplementary Material illustrates the relationship between the description of the inverted-waves image and the dress and jacket scores. We excluded "other" answers from the analyses and calculated a one-way ANOVA with the five descriptions of the ambiguous image as the factor (This ANOVA tests the difference between the first five bars in Fig. S6). The difference between the scores of the dress (Fig. S6.a) was close to significance ($F(4446) = 2.3$, $p = 0.06$). The three ANOVAS for the Jacket, the Dress2 and Dress3 (Fig. S6.b-d) were not even close to significance ($F(4446) = 0.44$–1.2, $p = 0.29$–0.75). There were also no significant effects in post hoc t-tests (see "Inverted Waves" in the Supplementary material).

3.2.7. Subjective grey and memory colour effect

To examine the relationship between perceived colour of DressO and subjective grey points, we compared dress scores in an independent *t*-test between observers who chose the grey and those who chose the bluish exemplar of the disk or banana. This was done with the answers to the disk and to the banana, separately. Fig. 8 and Fig. S7 illustrate the results for DressO in the main and in the preliminary experiment, and Table S7 reports detailed results from t-tests.

Scores for DressO were very similar between the observers who chose the grey and those who chose the bluish instance of the disk and the banana respectively. Scores for DressO did not differ between the groups for the disk and the banana in the preliminary experiment (both $p > 0.91$ cf. Fig. S7) and in the main experiment (both $p > 0.73$; cf. Fig. 8). There were no systematic relationships between the choices of grey for the disk and the banana and the scores for any of the three other colour-ambiguous images (for details see "Subjective grey in main experiment" in the Supplementary material).

The comparison between results for the disk and for the banana gives insight into effects that are specific to memory colours (Witzel, 2016). However, there were no effects for either kind of stimuli. These results undermine the idea that the perceived colours of DressO or of any other colour-ambiguous image are related to individual differences in subjective grey-points and in memory colour effects.

3.2.8. Colour naming

Detailed results on the relationships between measured category boundaries and reported colours for the colour-ambiguous images are provided in section "Colour naming" of the Supplementary Material. For DressO, the dress score was positively correlated with the brown-white boundary (Fig. 9.d and Fig. S8.i; $r(403) = 0.18$, $p = 0.0003$). This correlation indicates that observers whose category boundary was closer to white than to brown tended to report that the dress was white-gold. This is unexpected because neither the perceived colours of the dress lace nor those of the dress body vary along the brown-white dimensions.

Moreover, the scores of Dress2 were negatively correlated with the brown-yellow boundary (Fig. 9.c and Fig. S8.r; $r(402) = -0.17$, $p = 0.0006$) and the light green-blue boundary (gb2 in Fig. 9.e ($r(403) = -0.16$, $p = 0.001$). These correlations were still significant when applying a Bonferroni correction for five measurements of category boundaries (alpha = 0.01). These correlations make sense since they indicate that observers who generally tended to name comparatively many colours as green and brown instead of blue and yellow also tended to describe Dress2 as green/grey (body) and bronze/gold (lace) rather than blue and white. The same correlations as for Dress2 were found for Dress3 that is, with brown-yellow (Fig. 9.c and Fig. S8.w; $r(402) = -0.10$, $p = 0.04$) and light green-blue (Fig. 9.e and Fig. S8.y; $r(403) = -0.11$, $p = 0.03$). Since the content of this photo is the same as for Dress2 except for the white-point adjustment, it makes sense that it features similar correlations to Dress2. However, the correlations for Dress3 do not survive a Bonferroni correction. There were no other significant correlations, in particular none involving the Jacket.

We conducted further analyses (see section "Colour Naming" in the Supplementary Material), which revealed that differences across countries of origin, and hence across languages, might have distorted the relationship between reported dress colours and colour naming. If the individual differences in reported colours of DressO were due to differences in colour naming, then language-specific effects on naming should directly affect reported dress colours and hence increase correlations between dress colours and naming, especially with the blue-white and the brown-yellow boundary. However, this was not the case.

3.2.9. Relationship between tasks

Finally, we examined relationships between individual differences in the gloss ratings, the task on specular highlights and colour constancy, in the perception of the inverted-waves image and in the subjective grey-point. However, we did not find any relationship between those tasks. Details on these analyses are provided in the section "Relationships between tasks" in the Supplementary Material.

3.3. Discussion

3.3.1. How the dress phenomenon works

The results of the online experiments provided strong evidence that the perceived colours of the original DressO are related to the assumed illumination of the dress. These results replicate and extend earlier findings (Chetverikov & Ivanchei, 2016; Toscani, Gegenfurtner, & Doerschner, 2017; Wallisch, 2017; Witzel et al., 2017). In particular, the present findings reproduce the relationship between perceived dress colours and the assumed colour of the illumination along the blue-yellow dimension. In addition to that, the relationship between perceived dress colours and assumed illumination could even be shown with the small dataset in the preliminary experiment for brightness but not for colour estimations. This suggests that the assumed brightness of the illumination might be even more important in determining the perceived colours of the dress than the assumed colour of the illumination. This is in line with the observation that the perceived colours of the dress vary most strongly along the lightness dimension (Gegenfurtner et al., 2015). Different assumptions about the brightness of the illumination explain why observers perceive the dress in strongly different levels of lightness.

Among the questions about the source of the illumination, the belief that the dress is in the shadow is very clearly related to the perceived colours of the dress. Assumptions suggesting a strong direct illumination, such as a flash, overexposure and maybe direct daylight also shape the perception of the dress colours. These results show that observers who see the dress in a bright illumination (flash, overexposure) perceive its colours as blue and black. In contrast, those who consider it to be in the shadow perceive it as white and gold because they attribute the dark blue colour to the illumination rather than to the dress itself. This is in line with the idea that the differences in colour perception are related to a difference in whether observers attribute perceived features to the surface or to the illuminating light.

We also observed that the perceived colour of the dress is related to the gloss of the dress. Observers who saw more gloss in the lace and the bolero of the dress also tended to see it as white-gold rather than blue-black. Moreover, observers who believed the gloss and shininess of the dress was due to a strong light tended to see the dress as blue-black; those who thought gloss and shininess were due to the fabric tended to see the dress as white-gold. These observations are in line with the fact that gold implies the presence of glossiness, and with the overall idea that the perception of the dress is due to assumptions about the illumination.

The colour naming measurements indicated a relationship between the dress and the white-brown dimension (Fig. S8.i). However, the meaning of this relationship is not clear and further investigations are needed to clarify the precise role of colour naming for the reported colours of the dress. In any case, the relationship between colour naming and the perceived colours of the dress seems to be minor compared to the robust effect of assumptions about the illumination (see additional analyses in the section "Colour Naming" in the Supplementary Material).

Finally, the absence of any effects of subjective grey choices on DressO is in line with the results in the previous laboratory experiment (Witzel et al., 2017). This further supports the conclusion that there is at least no simple relationship between subjective grey points and the perceived colours of the dress.

3.3.2. Global perceptual strategies

With a few exceptions, the other kinds of individual differences we observed seem to be largely independent of the original photo of the dress and of each other. First of all, Jacket, Dress2 and Dress3 were strongly related to each other, but not to DressO. There were a few indications that the individual differences for the Jacket might be due to similar reasons as those for DressO. Most notably, there was a relationship between brightness estimations of the illumination and the perceived colours of the Jacket, suggesting that the perceived colours are modulated by implicit assumptions about the illumination, as was the case for DressO. This observation supports the idea that the assumed brightness of the illumination might play a more general role for colour-ambiguous images. However, the answers to the questions about the illumination in the photo of the Jacket did not support a relationship between assumed illumination and perceived colour of the Jacket (cf. Fig. 12.e-g). One reason for this might be that the questions about the illumination did not capture the relevant characteristics of the illumination in the photo of the Jacket.

Moreover, we found some evidence that gloss ratings were related between DressO and the jacket. However, the meaning of such relationships is unclear and the gloss ratings were not related to the perceived colours of the Jacket. Furthermore, we found that the reported colours of Dress2 (Fig. 4.g-i) were related to individual differences in naming brown and yellow, and (light) green and blue colours (Fig. 9.c,e and Fig. S8.r,t). However, evidence for the role of colour naming was complicated by the effect of language, and hence it is as yet unclear how much Jacket, Dress2 and Dress3 really depend on colour naming. Overall our findings rather suggest that individual differences in perceived colours for the Jacket, Dress2 and Dress3 are different phenomena than those for DressO.

Our results confirm the observation of Häkkinen and Gröhn (2016) that the inverted-waves photo is perceived very differently by different observers (Fig. 7). Inverting the photo of the waves has a strong effect on perception that is likely to be related to perceptual inferences of three-dimensional shape from two-dimensional cues on shading (Häkkinen & Gröhn, 2016; Ramachandran, 1988). However, according to our data there seemed not to be a relationship with the perception of DressO (Fig. S6.a) and the other phenomena.

Like Lee and Smithson (2016) we found average performance in the colour constancy task was above chance, indicating that there is, on average, a contribution of gloss to colour constancy. This is in line with studies that found that observers use gloss as a cue about the illumination (Snyder, Doerschner, & Maloney, 2005) and that colour constancy is significantly higher in the presence of gloss (e.g. Granzier, Vergne, & Gegenfurtner, 2014). At the same time, many observers had difficulties in using gloss as a cue for colour constancy. Moreover, the variation across observers in this task was not related to other kinds of individual variation. We think that the individual differences in that task depend on whether observers understood the relationship between the highlight and the correct response in the task. If observers understood that looking at the gloss completely solves the task, their performance should be close perfect. We had asked participants in a pilot trial to write us their impression about this task. When asked how he achieved 100% correct performance in this task, a participant replied: "I see, no surprise, just got the trick. Haha". The point was that he had figured out the logical relationship between highlight and illumination changes and hence could answer correctly based on this knowledge. We speculate that this might also be the reason for the individual differences in Lee and Smithson

(2016)'s original study, because observers with high performance tended to be experienced (the two authors and three observers with training in colour vision), while those with low performance tended to be inexperienced with this kind of task (two completely naïve observers and two observers with training in colour vision). If this is so, it is no wonder that individual variation in this task is not related to the other tasks because responses in those other tasks depend on subjective appearance and perception, not on figuring out a logical relationship. This idea is illustrated by the observation that perception in those other tasks, e.g. of DressO or the inverted waves, cannot be changed by simply informing observers about the cues in the images.

In line with previous experiments (Hansmann-Roth et al., 2015; Hansmann-Roth et al., 2017), we also found individual differences in gloss ratings. In our results, gloss ratings were not bimodally distributed, but spread across the whole rating scale. According to these results, gloss perception varies continuously across individuals rather than being separated into two distinct groups of observers. However, there was little evidence for a relationship between those images with ambiguous gloss and the perception of DressO or other phenomena. Overall these results do not support the idea that individual differences in gloss perception are related to other kinds of individual differences. Besides the idea of general differences in perceptual strategies, there are several other possible reasons why individual differences in gloss perception might occur with two-dimensional images, but not with real objects. In particular, two-dimensional images lack high dynamic range, there is no effect of movement on the location of highlights, and there is no binocular-disparity, which all play a role in the perception of gloss (Chadwick & Kentridge, 2015 for review; Doerschner et al., 2011; Doerschner, Maloney, & Boyaci, 2010; Kitazaki, Kobiki, & Maloney, 2008; Wendt, Faul, & Mausfeld, 2008). Dynamic range might be a good candidate since it determines the range of contrasts and contrasts play an important role in gloss perception (Chadwick & Kentridge, 2015 for review; Wiebel, Toscani, & Gegenfurtner, 2015). At the same time the highlights in the images that yielded individual differences in our study have either particularly low (white and glossy) or particularly high (black and glossy) contrast. It might be that observers discount and compensate for the absence of high dynamic range to different degrees. This would be rather specific to those images, and hence it would explain why gloss ratings were not related to other tasks.

In sum, our findings from the online study confirm that there are individual differences in all the phenomena and tasks we investigated. However, these differences are largely unrelated to each other. According to the experiment on colour constancy, the dress (i.e. DressO) is related to individual differences in colour constancy. There was some evidence that this relationship between the dress and colour constancy might be specific for surface colours and illuminations along the daylight axis; but at the current state we cannot conclude about the precise nature of the relationship between colour constancy and the dress with certainty. At the same time, the results from the online experiment suggest that the phenomena and tasks on the perception of surface properties investigated in the online experiment were quite different from the phenomenon of the dress. Other factors seem to intervene in the determination of individual differences in those tasks.

3.3.3. Limitations of online surveys

The present results also allow for evaluating the use of online surveys for experiments on perception. In a previous study (Witzel, 2016) memory colour effects which involve subtle changes in colour appearance were reproduced with data from online surveys. The present online experiments provide another case where results from the laboratory could be reproduced through online surveys.

The agreement between online and laboratory experiments is not trivial because the most relevant stimulus dimensions, such as colour and gloss, are completely uncalibrated in online studies because they depend on the devices used by participants. Hence, particular care needs to be taken to ensure that the lack of stimulus control does not produce spurious findings. The comparison with laboratory experiments, in which devices are calibrated and stimuli are carefully controlled, allows for cross-validating results from online studies. The success of doing so in the present study provides further support for the idea that online studies may be a meaningful extension of the experimental repertoire. Although online studies definitely have limitations as controlled experiments they also have advantages that make them more useful for certain purposes than experimental studies in the laboratory (Witzel, 2016; Woods, Velasco, Levitan, Wan, & Spence, 2015).

3.3.4. Pre-constancy vision in infants

Our results revealed that the images that were taken as examples for matte stimuli by Yang et al. (2015) are actually unambiguously glossy. These supposedly matte images were generated by increasing the specular reflection blur and manually removing the highlight region (see Yang et al., 2015, Supplementary Material). However, the interreflections within the environment are still visible and reflected on the material of the object and the supposedly matte images therefore appeared glossy to almost all of our observers.

This finding is interesting because it suggests that the difference the authors found between infants and adults are not related to the perception of gloss, contrary to what was claimed in their study (Yang et al., 2015). Their results showed that in contrast to older children three to four month old infants show higher novelty preferences when the direction of the illumination (the "light-field") changes than when the surface changes from glossy to matte (cf. Fig. 2D in Yang et al., 2015). According to the authors, these results show "that before developing perceptual constancy, 3- to 4-month-old infants have a striking ability to discriminate slight image changes due to illumination that are not salient for adults" and that "[t]hese young infants lose this ability after 5 months of age and then develop an ability to perceive distal surface properties (glossy or matte) at 7–8 months of age." (abstract in Yang et al., 2015, p. 3209). However, since all of their stimuli are clearly glossy, the distinction between matte and glossy surfaces cannot be the relevant factor that distinguishes young infants' responses to those images from the responses of older infants and adults. This idea is further supported by another finding reported in that study (Yang et al., 2015): They created images with scrambled textures, in which three-dimensional shape and gloss were completely removed. Although none of those images looked glossy, Yang et al. (2015) still observed the difference in novelty preferences between 3 and 4 year old and older infants.

In sum, the effect observed by Yang et al. (2015) occurs for stimulus images that are all glossy and for stimulus images that are scrambled and not glossy at all. Hence, this effect does not depend on gloss perception and there is no reason to believe that it is related to surface constancy at all. It would be important to know where the effect observed by Yang et al. (2015) actually comes from.

4. Conclusion

Taken together, the results of this study further clarify why different observers perceive the colours of the dress (DressO) differently. The findings from the online study show that the inter-individual differences in the perception of the colour and gloss of the dress are related to whether colour and gloss are attributed to the surface or to the light that is illuminating the dress. These

results replicate and extend previous findings (Chetverikov & Ivanchei, 2016; Toscani et al., 2017; Wallisch, 2017; Witzel et al., 2017).

The present findings also provide first insights into the relationship between individual differences in different kinds of tasks involving the perception of surface properties. First of all, the perception of the dress was strongly related to a typical colour constancy task. This finding shows that the individual differences in the perception of the dress are not limited to that particular photo of the dress, but are related to more general aspects of colour constancy (see also Weiss, Gegenfurtner, & Witzel, accepted). Moreover, this observation also highlights the fact that individual differences in colour constancy are not noise, but rather systematic and reliable (Witzel, van Alphen, et al., 2016). At the same time, the individual differences investigated in the online experiment were largely independent of the dress and of each other. This was also true for individual differences in the description of other colour-ambiguous images, such as the Jacket (poppunkblogger, 2016). According to these results inter-individual differences arise for various reasons and may be unrelated across the different perceptual domains. Taken together, our findings suggest that the individual differences observed for the dress are related to specific aspects of colour constancy, but not to more general perceptual strategies that encompass other domains beyond colour constancy, such as gloss and shape from shading. The specificity of the phenomenon of the dress might be one of the reasons why the dress has yielded an unprecedented interest in the social media and the broader public. Yet, it is unclear what specific aspects of colour constancy are related to the dress and why other situations involving colour constancy do not feature the striking individual differences observed for the dress.

In addition to these main findings, our results also inform previous findings on gloss perception in infants (Yang et al., 2015). Moreover, the successful replication of experimental results in the online study supports the idea that online experiments may be useful for investigations on visual perception despite the absence of display calibration and proper colour rendering.

Acknowledgments

We are grateful to Jiale Yang and his colleagues for providing the two images from their infant study; to Ying Chen for posing for Dress2 and Dress3; to Carlijn van Alphen for help with data collection; and to Alice Skelton, Chris Racey, John Maule and Bart Anderson for helpful discussion. This work was supported by ERC Advanced grant "FEEL" No 323674 to J. Kevin O'Regan and by grant "Cardinal Mechanisms of Perception" No SFB TRR 135 from the Deutsche Forschungsgemeinschaft. SHR was supported by the EU Marie Curie Initial Training Network "PRISM" (FP7-PEOPLE-2012-ITN, Grant Agreement: 316746).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.visres.2017.07.015.

References

Adobe Systems Incorporated. (2008). Adobe Photoshop Elements 7 (Version 7.0).

Arend, L. E., Reeves, A., Schirillo, J., & Goldstein, R. (1991). Simultaneous color constancy: Paper with diverse Munsell values. *Journal of the Optical Society of America A, 8*(4), 661–672.

Bach, M. (2015). Aufregung um die Farbe eines Kleides. *Der Ophthalmologe, 112*(6), 512–516. http://dx.doi.org/10.1007/s00347-015-0064-0.

Bosten, J. M., Beer, R. D., & MacLeod, D. I. (2015). What is white? *Journal of Vision, 15*(16), 5. http://dx.doi.org/10.1167/15.16.5.

Brainard, D. H., & Hurlbert, A. C. (2015). Colour vision: Understanding #TheDress. *Current Biology, 25*(13), R551–R554. http://dx.doi.org/10.1016/j.cub.2015.05.020.

Chadwick, A. C., & Kentridge, R. W. (2015). The perception of gloss: a review. *Vision Research, 109*, 221–235. http://dx.doi.org/10.1016/j.visres.2014.10.026.

Chauhan, T., Perales, E., Xiao, K., Hird, E., Karatzas, D., & Wuerger, S. (2014). The achromatic locus: Effect of navigation direction in color space. *Journal of Vision, 14*(1). http://dx.doi.org/10.1167/14.1.25.

Chetverikov, A., & Ivanchei, I. (2016). Seeing "the Dress" in the Right Light: Perceived Colors and Inferred Light Sources. *Perception*. http://dx.doi.org/10.1177/0301006616643664.

CNRS. (2015). Relais d'Information sur les Sciences de la Cognition (RISC). from Centre National de la Recherche Scientifique (CNRS) http://www.risc.cnrs.fr/.

Debevec, P. (1998). Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. Paper presented at the Proceedings of the 25th annual conference on Computer graphics and interactive techniques.

Doerschner, K., Fleming, R. W., Yilmaz, O., Schrater, P. R., Hartung, B., & Kersten, D. (2011). Visual motion and the perception of surface material. *Current Biology, 21*(23), 2010–2016. http://dx.doi.org/10.1016/j.cub.2011.10.036.

Doerschner, K., Maloney, L. T., & Boyaci, H. (2010). Perceived glossiness in high dynamic range scenes. *J Vis, 10*(9), 11.

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. http://dx.doi.org/10.1167/17.3.15.

Foster, D. H. (2011). Color constancy. *Vision Research, 51*(3), 647–700. http://dx.doi.org/10.1016/j.visres.2010.09.006.

Gegenfurtner, K. R., Bloj, M., & Toscani, M. (2015). The many colours of 'the dress'. *Current Biology, 25*(13), R543–R544. http://dx.doi.org/10.1016/j.cub.2015.04.043.

Granzier, J. J. M., Brenner, E., & Smeets, J. B. (2009). Reliable identification by color under natural conditions. *Journal of Vision, 9*(1). 39 31-38. doi:10.1167/9.1.39; /9/1/39/ [pii].

Granzier, J. J. M., & Gegenfurtner, K. R. (2012). Effects of memory colour on colour constancy for unknown coloured objects. *i-Perception, 3*(3), 190–215. http://dx.doi.org/10.1068/i0461.

Granzier, J. J. M., Vergne, R., & Gegenfurtner, K. R. (2014). The effects of surface gloss and roughness on color constancy for real 3-D objects. *Journal of Vision, 14*(2). http://dx.doi.org/10.1167/14.2.16.

Häkkinen, J., & Gröhn, L. (2016). Turning water into rock: The inverted waves effect. *i-Perception, 7*(1). http://dx.doi.org/10.1177/2041669515627951.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience, 9*(11), 1367–1368. nn1794 [pii], DOI: 10.1038/nn1794.

Hansmann-Roth, S., Pont, S. C., & Mamassian, P. (2015). Gloss perception of photographs and real multi-material objects. *Perception, 40*(1 Abstract Supplement), 53.

Hansmann-Roth, S., Pont, S. C., & Mamassian, P. (2017). Contextual effects on real bicolored glossy surfaces. *Journal of Vision, 17*(2), 17. http://dx.doi.org/10.1167/17.2.17.

Hesslinger, V. M., & Carbon, C.-C. (2016). #TheDress: The role of illumination information and individual differences in the psychophysics of perceiving white-blue ambiguities. *i-Perception, 7*(2). http://dx.doi.org/10.1177/2041669516645592.

Hurlbert, A., Aston, S., & Pearce, B. (2016). Is that really# thedress? Individual variations in colour constancy for real illuminations and objects. *Journal of Vision, 16*(12), 743.

Ishihara, S. (2004). *Ishihara's tests for colour deficiency*. Tokyo, Japan: Kanehara Trading Inc.

Isis Software Incubator. (2015). Prolific Academic. from Isis Innovation Ltd https://prolificacademic.co.uk/.

Karlsson, B. S. A., & Allwood, C. M. (2016). What is the correct answer about The Dress' colors? Investigating the relation between optimism, previous experience, and answerability. *Frontiers in Psychology, 7*(1808). http://dx.doi.org/10.3389/fpsyg.2016.01808.

Kitazaki, M., Kobiki, H., & Maloney, L. T. (2008). Effect of pictorial depth cues, binocular disparity cues and motion parallax depth cues on lightness perception in three-dimensional virtual scenes. *PLoS ONE, 3*(9), e3177. http://dx.doi.org/10.1371/journal.pone.0003177.

Kohonen, O., Parkkinen, J., & Jaaskelainen, T. (2006). Databases for spectral color science. *Color Research and Application, 31*(5), 381–390. http://dx.doi.org/10.1002/Col.20244.

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. http://dx.doi.org/10.1016/j.cub.2015.04.053.

Lee, R. J., & Smithson, H. E. (2016). Low levels of specularity support operational color constancy, particularly when surface and illumination geometry can be inferred. Journal of the Optical Society of America. A, Optics, image science, and vision, 33(3), A306–A318. doi:10.1364/JOSAA.33.00A306.

Lindsey, D. T., & Brown, A. M. (2009). World Color Survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences USA*. 0910981106 [pii] 10.1073/pnas.0910981106.

Macknik, S. L., Martinez-Conde, S., & Conway, B. R. (2015). Unraveling "The Dress". *Scientific American Mind, 26*(4), 19–21. http://dx.doi.org/10.1038/scientificamericanmind0715-19.

Mahroo, O. A., Williams, K. M., Hossain, I. T., Yonova-Doing, E., Kozareva, D., Yusuf, A., & Hammond, C. J. (2017). Do twins share the same dress code? Quantifying relative genetic and environmental contributions to subjective perceptions of "the dress" in a classical twin study. *Journal of Vision, 17*(1), 29. http://dx.doi.org/10.1167/17.1.29.

Moccia, M., Lavorgna, L., Lanzillo, R., Brescia Morra, V., Tedeschi, G., & Bonavita, S. (2016). The Dress: Transforming a web viral event into a scientific survey. *Multiple Sclerosis and Related Disorders, 7*, 41–46. http://dx.doi.org/10.1016/j.msard.2016.03.001.

Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2008). Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *Journal of Vision, 8*(5), 1–16.

Olkkonen, M., Witzel, C., Hansen, T., & Gegenfurtner, K. R. (2010). Categorical color constancy for real surfaces. *Journal of Vision, 10*(9), 1–22. 10.9.16 [pii] 10.1167/10.9.16.

Parkkinen, J. P. S., Hallikainen, J., & Jaaskelainen, T. (1989). Characteristic spectra of munsell colors. *Journal of the Optical Society of America a-Optics Image Science and Vision, 6*(2), 318–322. http://dx.doi.org/10.1364/Josaa.6.000318.

Pernet, C. R., Wilcox, R., & Rousselet, G. A. (2012). Robust correlation analyses: False positive and power validation using a new open source matlab toolbox. *Frontiers in Psychology, 3*, 606. http://dx.doi.org/10.3389/fpsyg.2012.00606.

poppunkblogger. (2016). One year after #TheDress. Retrieved from http://wondirwin.tumblr.com/post/140009149046/poppunkblogger-dammitmichael.

Rabin, J., Houser, B., Talbert, C., & Patel, R. (2016). Blue-black or white-gold? Early stage processing and the color of 'the dress'. *PLoS ONE, 11*(8), e0161090. http://dx.doi.org/10.1371/journal.pone.0161090.

Radonjic, A., & Brainard, D. H. (2016). The nature of instructional effects in color constancy. *Journal of Experimental Psychology: Human Perception & Performance, 42*(6), 847–865. http://dx.doi.org/10.1037/xhp0000184.

Ramachandran, V. S. (1988). Perception of shape from shading. *Nature, 331*(6152), 163–166. http://dx.doi.org/10.1038/331163a0.

Schlaffke, L., Golisch, A., Haag, L. M., Lenz, M., Heba, S., Lissek, S., & Tegenthoff, M. (2015). The brain's dress code: How The Dress allows to decode the neuronal pathway of an optical illusion. *Cortex, 73*, 271–275. http://dx.doi.org/10.1016/j.cortex.2015.08.017.

Snyder, J. L., Doerschner, K., & Maloney, L. T. (2005). Illumination estimation in three-dimensional scenes with and without specular cues. *Journal of Vision, 5*(10), 863–877. http://dx.doi.org/10.1167/5.10.8 /5/10/8/ [pii].

Swiked. (2015). guys-please-help-me-is-this-dress-white-and. Retrieved from http://swiked.tumblr.com/post/112073818575/guys-please-help-me-is-this-dress-white-and.

Toscani, M., Dörschner, K., & Gegenfurtner, K. (2016). Probing the illumination on #The Dress. *Journal of Vision, 16*(12), 633. http://dx.doi.org/10.1167/16.12.633.

Toscani, M., Gegenfurtner, K. R., & Doerschner, K. (2017). Differences in illumination estimation in #thedress. *Journal of Vision, 17*(1), 22. http://dx.doi.org/10.1167/17.1.22.

Vemuri, K., Bisla, K., Mulpuru, S., & Varadharajan, S. (2016). Do normal pupil diameter differences in the population underlie the color selection of #thedress? *Journal of the Optical Society of America A, 33*(3), A137–A142. http://dx.doi.org/10.1364/JOSAA.33.00A137.

Wallisch, P. (2017). Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain: "The dress"Wallisch. *Journal of Vision, 17*(4), 5. http://dx.doi.org/10.1167/17.4.5.

Walter, B., Marschner, S. R., Li, H., & Torrance, K. E. (2007). Microfacet models for refraction through rough surfaces. Paper presented at the Proceedings of the 18th Eurographics conference on Rendering Techniques, Grenoble, France.

Webster, M. A., & Kay, P. (2007). Individual and population differences in focal colors. In R. E. MacLaury, G. V. Paramei, & D. Dedrick (Eds.), *Anthropology of color*. Amsterdam: John Benjamins.

Weiss, D., Gegenfurtner, K. R., & Witzel, C. (accepted). #thedress reveals general "chromotypes" in colour constancy Paper presented at the 40th European Conference on Visual Perception (ECVP), Berlin, Germany.

Weiss, D., Witzel, C., & Gegenfurtner, K. R. (under review). Determinants of colour constancy and the blue bias.

Wendt, G., Faul, F., & Mausfeld, R. (2008). Highlight disparity contributes to the authenticity and strength of perceived glossiness. *Journal of Vision, 8*(1), 14. http://dx.doi.org/10.1167/8.1.14.

Werner, A., & Schmidt, A. (2016). The #Dress phenomenon an empirical investigation into the role of the background. *Journal of Vision, 16*(12), 742. http://dx.doi.org/10.1167/16.12.742.

Wexler, M., Duyck, M., & Mamassian, P. (2015). Persistent states in vision break universality and time invariance. *Proceedings of the National Academy of Sciences, 112*(48), 14990–14995. http://dx.doi.org/10.1073/pnas.1508847112.

Wiebel, C. B., Toscani, M., & Gegenfurtner, K. R. (2015). Statistical correlates of perceived gloss in natural images. *Vision Research, 115*(Pt B), 175–187. http://dx.doi.org/10.1016/j.visres.2015.04.010.

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. http://dx.doi.org/10.1016/j.cub.2015.05.004.

Witzel, C. (2012). Colours' appearance in the light of language and experience. (Dr. rer. nat. Dissertation), Justus-Liebig-Universität, Gießen.

Witzel, C. (2015). The Dress: Why do different observers see extremely different colours in the same photo? Retrieved from http://lpp.psycho.univ-paris5.fr/feel/?page_id=929.

Witzel, C. (2016). An easy way to show memory color effects. *i-Perception, 7*(5), 1–11. http://dx.doi.org/10.1177/2041669516663751.

Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of Vision, 13*(7). http://dx.doi.org/10.1167/13.7.1.

Witzel, C., Hansmann-Roth, S., & O'Regan, J. K. (2016). Individual differences in the perception of surface properties. *Perception, 45*(S2), 189.

Witzel, C., O'Regan, J. K., & Rothen, N. (2016). *Synaesthetic colour experiences are perceptually real*. Paper presented at the Synaesthesia and Cross-Modal Perception, Dublin, Ireland.

Witzel, C., Olkkonen, M., & Gegenfurtner, K. R. (2016). Memory colours affect colour appearance. *Behavioral and Brain Sciences, 39*, 51–52. http://dx.doi.org/10.1017/S0140525X15002587.

Witzel, C., Racey, C., & O'Regan, J. K. (2016). Perceived colors of the color-switching dress depend on implicit assumptions about the illumination. *Journal of Vision, 16*(12), 223. http://dx.doi.org/10.1167/16.12.223.

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. http://dx.doi.org/10.1167/17.2.1.

Witzel, C., Valkova, H., Hansen, T., & Gegenfurtner, K. R. (2011). Object knowledge modulates colour appearance. *i-Perception, 2*(1), 13–49. http://dx.doi.org/10.1068/i0396.

Witzel, C., van Alphen, C., Godau, C., & O'Regan, J. K. (2016). Uncertainty of sensory signal explains variation of color constancy. *Journal of Vision, 16*(15), 8. http://dx.doi.org/10.1167/16.15.8.

Witzel, C., Wuerger, S., & Hurlbert, A. (2016). Variation of subjective white-points along the daylight axis and the colour of the dress. *Journal of Vision, 16*(12), 744. http://dx.doi.org/10.1167/16.12.744.

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: a tutorial review. *PeerJ, 3*, e1058. http://dx.doi.org/10.7717/peerj.1058.

Wuerger, S. M., Hurlbert, A. C., & Witzel, C. (2015). Variation of subjective white-points along the daylight axis and the colour of the dress. *Perception, 44*(S1), 153.

Yang, J., Kanazawa, S., Yamaguchi, M. K., & Motoyoshi, I. (2015). Pre-constancy vision in infants. *Current Biology, 25*(24), 3209–3212. http://dx.doi.org/10.1016/j.cub.2015.10.053.