Vision Research 151 (2018) 117–134

Contents lists available at ScienceDirect

Vision Research

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

# An evaluation of different measures of color saturation

Florian Schiller [*], Matteo Valsecchi, Karl R. Gegenfurtner

*Department of Psychology, Justus Liebig University Giessen, Giessen, Germany*

ARTICLE INFO

*Article history:*
Received 19 December 2016
Received in revised form 25 March 2017
Accepted 3 April 2017
Available online 9 June 2017

Number of Reviews = 2

*Keywords:*
Color saturation
Saturation measure
Luminance contrast
Brightness
Visual cortex

ABSTRACT

We investigated how well seven saturation measures defined in CIECAM02, HSV, DKL, LAB, LUV, and CIE 1931 xyY color spaces correspond to human perception of saturation. We used a paradigm that allowed us to measure the perceived saturation of several standard color stimuli in many different directions of color space. We implemented this paradigm at different levels of luminance and varied background luminance relative to the luminance of our color stimuli in order to ensure the generality of our approach. We found that varying background luminance changed the relative saturation of the standard colors. Raising the overall luminance level did not have such an effect. We compared the results of our measurements to the predictions of the seven saturation measures. All of the measures could predict our observers' judgments of saturation reasonably well. The measures that are based on measurements of discrimination thresholds (LUV, LAB, CIECAM02) performed best on average. However, some of the perceptual effects induced by changing background luminance could not be predicted by any measure.

## 1. Introduction

The results of Maxwell (1857) and Helmholtz (1852), ingeniously summarized and extended by Schrödinger (1920), show that the color of any light stimulus can be matched by a weighted combination of three arbitrary primary colors. The arbitrariness of the three primary color dimensions is in stark contrast to our phenomenological experience of color, where we usually use color as a synonym for "hue" that is then further specified by an intensity (luminance, brightness or lightness) and its saturation. In the past, the hue dimension has been thoroughly investigated and there has been even more work on luminance (cf. Wyszecki & Stiles, 1982). The third dimension, saturation, has been thoroughly neglected. This is not to say, however, that saturation as a perceptual dimension has remained entirely unexplored.

It is well established that perceived saturation is a function of colorimetric purity and dominant wavelength. The colorimetric purity of a color c is the amount of spectral light of the color's dominant wavelength relative to the amount of white light that is required to produce c (cf. Wyszecki & Stiles, 1982; Hunt & Pointer, 2011, for instance). The influence of colorimetric purity (or correlates thereof) and dominant wavelength was not only examined in measurements of just noticeable differences (JNDs) (Aubert, 1865; Jones & Lowry, 1926; Kaiser, Comerford, & Bodinger, 1976), but also in color matching tasks (Witzel & Franklin, 2014; Zemach, Chang, & Teller, 2007), in scaling tasks (Indow, 1978; Indow & Stevens, 1966), and a forced choice paradigm (Switkes, 2008; cf. also Switkes & Crognale, 1999) where cone contrasts were determined for two chromatic gratings that were perceived as having equal contrast salience, which seems to be roughly equivalent to saturation as investigated here.

Experiments by Hunt (1950, 1952) indicate that a color's colorimetric purity and dominant wavelength are not the only determinants of its perceived saturation. Using a haploscope, Hunt (1950, 1952) presented a standard color on neutral background to one eye of his observers, and a comparison color on a neutral background to the other eye. Observers were asked to match the color of the comparison to the standard color in purity and luminance. Hunt found that if the luminance of the standard color and of its background was increased, then the comparison color needed to be set to a higher purity in order to match the standard. This result is known as the "Hunt effect" and often summarized by saying that a stimulus appears more colorful as its luminance is increased (Fairchild, 1998, pp. 144–145). The Hunt effect has been replicated using different paradigms such as the short-term memory paradigm reported in Pitt and Winter (1974) where observers matched a comparison in a dark surround to a standard whose luminance could be varied in a bright surround as they looked at each of the stimuli in turn with both eyes. Pitt and Winter (1974) found that the perceived saturation of the standard increased with its luminance and the luminance of its surround. Using a scaling task where observers had to estimate a test field's saturation after

[*] Corresponding author at: Department of Psychology, Justus Liebig University Giessen, Otto-Behagel-Str. 10F, D-35394 Giessen, Germany.
*E-mail address:* Florian.Schiller@psychol.uni-giessen.de (F. Schiller).

http://dx.doi.org/10.1016/j.visres.2017.04.012
0042-6989/© 2017 Elsevier Ltd. All rights reserved.

seeing an adaptation field, Jacobs (1967) showed that chromatic adaptation can affect perceived saturation as well. Hence, perceived saturation of a color seems to depend on the color's purity, its dominant wavelength, its luminance, the luminance of its surround, and chromatic adaptation.

Breneman (1977) reports that the Hunt effect can be diminished considerably by controlling for brightness and brightness contrast, which suggests that a color's perceived saturation may be influenced by the difference between the color's luminance and the luminance of its surround. This has been confirmed by Faul, Ekroll, and Wendt (2008) who presented a standard color patch on neutral background on one side of a computer screen to their observers. On the other side, they presented a comparison patch on a neutral background. This background could be more or less luminant than the background of the standard patch. Observers were asked to match the comparison in purity and luminance to the standard. Faul et al. (2008) found that the standard was perceived as more saturated when luminance contrast between the comparison patch and its surround was decreased. This result was replicated in two further paradigms. Xing et al. (2015) presented a color patch on an achromatic surround to their observers. The luminance of the patch was held constant while the luminance of the surround was varied systematically. The observers had to estimate the saturation of the patch. Bimler, Paramei, and Izmailov (2006, 2009) showed spectral colors of different wavelengths on a neutral background to their observers. The luminance of the spectral color was held constant while the lumiance of the background was varied. The observers were asked to name the spectral color. Using multidimensional scaling, Bimler et al. were able to determine the influence of changing luminance contrast on perceived saturation. Like Xing et al. (2015) they found that perceived saturation is decreased by increasing luminance contrast. So, to predict the perceived saturation of a color various variables have to be taken into account, such as its colorimetric purity, dominant wavelength, luminance, the luminance and chromaticity of its surround, and the corresponding luminance contrast.

Qualitative measures of saturation aim to predict perceived saturation. However, most of them neglect a considerable number of the saturation-relevant variables that have been discussed above. This may be related to the fact that, up to this date, there is no saturation measure based on direct measurements of perceived saturation. For instance, in CIE 1931 space, the saturation of a color c can be defined merely as the ratio of c's distance to the white point and the distance of c's dominant wavelength to the white point (this ratio is also called "excitation purity", cf. Oleari, 2016, p. 148). We will henceforth call this the "CIE measure". A modification of this measure that takes into account more than one point on the spectral locus was devised by Koenderink (2010) and will be called "KOE measure" here. CIE 1931 space itself is based on color matching functions that were measured in psychophysical experiments. Hence, the CIE 1931 space and the two measures defined in it are not based on measurements of saturation. To a lesser degree, the same applies to the two successors of the CIE 1931 space, LAB and LUV, which were suggested by the CIE as approximately perceptually uniform color spaces. Both color spaces are based on measurements of discrimination thresholds (Ohta & Robertson, 2005; Schanda, 2007, p. 58), such as those by MacAdam (1942).

However, discrimination thresholds may not accurately estimate perceived saturation. In LAB and LUV, saturation can be defined as a color's distance from the white point divided by its lightness. Hence, in these color spaces, the saturation measures are not based on direct measurements of saturation either. One of the latest color spaces suggested by the CIE is the CIECAM02 color space. In CIECAM02 space, saturation is defined as ratio of colorfulness to brightness (Luo & Li, 2013). Again, no direct measurements of perceived saturation have been used to create CIECAM02. The same applies to DKL color space, which is a device dependent color opponent space introduced by Krauskopf and colleagues (Derrington, Krauskopf, & Lennie, 1984; Krauskopf, Williams, & Heeley, 1982; see Hansen & Gegenfurtner, 2013). DKL color space is based on psychophysical and neurophysiological measurements of color opponent channels. Saturation can be defined as distance from the white point divided by luminance in this space. Similarly, the device dependent HSV (Hue, Saturation, Value) color space strives to capture the saturation dimension, as specified by its name, but is also not based on empirical measurements. HSV color space is simple transformation of RGB color space into cylindrical coordinates, where saturation is given by the radius.

All device independent measures mentioned above can take chromatic adaptation into account, although the CIE and KOE measures originally were not intended to do so (cf. Schiller & Gegenfurtner, 2016, Table 1). The numerical value that the CIECAM02 and the LAB measure assign to a color can furthermore be influenced by the color's luminance and the luminance of its surround. The magnitude of these changes in saturation can be substantial. For instance, changing the background from the CIE 1931 xyY coordinates $bg_1 = (0.331, 0.339, 40)$ to $bg_2 = (0.331, 0.339, 90)$ increases the saturation of the color $c_1 = (0.382, 0.285, 5)$ from $S_{LAB}(c_1, bg_1) = 0.68$ to $S_{LAB}(c_1, bg_2) = 0.77$ according to the LAB measure. Increasing the luminance of the color $c_1$ to $c_2 = (0.382, 0.285, 30)$ while background luminance stays at $Y = 40$ decreases saturation to $S_{LAB}(c_2, bg_1) = 0.58$. Such changes are not predicted by the other measures. The numerical value that is assigned to a color saturation by the CIE, KOE, HSV and the DKL measure does not change as the color's luminance or the luminance of its surround is changed.

Formal definitions of the seven measures mentioned above are provided in Schiller and Gegenfurtner (2016) who show that no pair of these measures is ordinally equivalent. That is, it is possible to find two colors $c_m$ and $c_n$ for each pair of measures $S_i$ and $S_j$ such

**Table 1**
Endpoints of comparison directions used in Experiment 1, 2, and 3.

| Direction | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Experiment 1 (patch luminance $Y = 30$ cd/m$^2$; background either 10 or 45 cd/m$^2$) | | | | | | | | | | |
| x | 0.252 | 0.181 | 0.187 | 0.194 | 0.212 | 0.419 | 0.506 | 0.597 | 0.483 | 0.381 |
| y | 0.190 | 0.256 | 0.324 | 0.394 | 0.645 | 0.492 | 0.425 | 0.355 | 0.269 | 0.209 |
| Experiment 2 (patch luminance $Y = 70$ cd/m$^2$; background either 50 or 120 cd/m$^2$) | | | | | | | | | | |
| x | 0.262 | 0.157 | 0.162 | 0.168 | 0.188 | 0.437 | 0.536 | 0.509 | 0.447 | 0.376 |
| y | 0.208 | 0.242 | 0.321 | 0.405 | 0.639 | 0.523 | 0.440 | 0.349 | 0.286 | 0.221 |
| Experiment 3 (patch luminance $Y = 50$ cd/m$^2$; background either 30, 40, 50, 60, or 70 cd/m$^2$) | | | | | | | | | | |
| x | – | 0.189 | – | – | 0.212 | – | – | 0.546 | – | – |
| y | – | 0.260 | – | – | 0.645 | – | – | 0.352 | – | – |

that $S_i(c_m) > S_i(c_n)$ while $S_j(c_m) \leq S_j(c_n)$. This raises the question as to which of the saturation measures agrees best with how humans perceive color saturation.

So far, this question has been addressed by Kim, Weyrich, and Kautz (2009), Cao et al. (2014), and Schiller and Gegenfurtner (2016). Kim et al. (2009) simultaneously presented three colored discs in the center of a gray background. One of the discs was neutral and acted as reference white, the second disc was colored and acted as a reference colorfulness patch, and the third stimulus was the test stimulus. The test stimulus could have one of 40 different colors. Participants were asked, among other things, to estimate the colorfulness of the test stimulus with regard to the reference, while background luminance, peak luminance, or ambient luminance was varied during different blocks of the experiment. Kim et al. (2009, Fig. 8) report that the CIECAM02 measure performed better than the LAB measure in sixteen out of nineteen experimental blocks. Cao et al. (2014) presented one of 33 different Munsell atlas samples above four reference samples on a gray easel under controlled and constant lighting conditions to their observers. Observers were asked to assign a number between zero and infinity to the test sample under the assumption that the reference stimuli have a saturation of one. On average, the LAB and CIECAM02 measures were closest to the judgments of the observers, followed by the LUV measure. Schiller and Gegenfurtner (2016) presented 80 images of natural scenes to their observers and asked them to select the most saturated spot with a mouse cursor. These choices were compared to the predictions of the seven measures described above. Schiller and Gegenfurtner found that all of the measures were able to predict the observers' choices reasonably well considering that none of the measures took the structure of the scene into account. The measures that are defined in color spaces based on discrimination thresholds, namely CIECAM02, LUV, and LAB, performed best on average, together with the measure defined in DKL color space.

As Schiller and Gegenfurtner (2016) point out, one problem of testing measures of saturation by using natural scenes as stimuli is that color distributions of natural scenes exhibit a bias in the yellowish-bluish direction. Thus, measures that perform well in this direction have an advantage over measures that do not. A less biased approach is taken by Cao et al. who selected their 33 test stimuli from ten color directions that were evenly distributed across the Munsell color system. However, working with only 33 test stimuli may have resulted in data that is too coarse to differentiate between the measures if the differences among them are small but systematic. The same problem applies to the study conducted by Kim et al. (2009) who used 40 test stimuli.

## 2. Experiments

To avoid the shortcomings of earlier studies, we used a dense sampling of color space in a forced choice paradigm, similar to the one used by Switkes and Crognale (1999) and Switkes (2008). Observers had to decide which of two equiluminant color patches presented against a neutral background was the more saturated. One patch always had one of three standard colors while the color of the other patch was sampled in a continuous way from one of ten color directions. This allowed us to determine points of equal saturation for each standard and color direction. These points could be compared to the predictions of the seven saturation measures in a metric of JNDs. Since perceived saturation is a function of luminance contrast, we also varied the luminance of the background between experimental sessions. Thus, we were able to determine the relative perceived saturation of different equiluminant hues as a function of their common background luminance. We implemented this approach to measuring saturation in Experiment 1. Since perceived saturation is influenced by the overall level of luminance as well according to the Hunt effect, we added Experiment 2. Experiment 2 is identical to Experiment 1 with the only difference that everything which was shown on the screen (i.e. the patches and their common background) was shifted to a higher level of luminance such that the brightness of the patches remained about the same as in Experiment 1. Since we found in both experiments that the relative saturation of two of the three standards reversed as background luminance was varied, we added Experiment 3. In Experiment 3, we varied background luminance of the two patches at five different levels to examine whether the reversal follows a linear course.

### 2.1. General methods

#### 2.1.1. Apparatus

The experimental setup of all experiments consisted of a monitor, a computer to control stimulus generation and presentation, a USB keyboard to record observers' responses, and chin rest.

In Experiment 1 and 3, we used a 22″ Eizo CG223W 10-bit LCD monitor (Eizo Nanao Corporation, Hakusan, Ishikawa, Japan) and a Dell Precision 380 computer (Dell Inc., Round Rock, Texas, USA). The chin rest held viewing distance constant at a distance of 0.4 m. The effective screen size of the monitor was 1680 × 1050 pixels, which corresponded to 0.474 × 0.296 m or 60° × 40° viewing angle. Color measurements at maximum power revealed the CIE xyY 1931 coordinates R = (0.6562, 0.3277, 35.68), G = (0.2137, 0.6831, 70.92), B = (0.1509 0.0698 8.958) for the red, green, and blue channel of the monitor, respectively. This results in the xy coordinates x = 0.331 and y = 0.339 for the monitor white. The Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and Matlab versions 2014a (The Mathworks Inc., Natick, MA, USA) were used for stimulus generation and presentation.

In Experiment 2, we used a different monitor to have a gamut at our disposal that was about as wide at higher luminance levels as the one we worked with in the other two experiments. The monitor was a 24 5/8″ PVM-2541 Sony (Sony Corporation, Minato, Tokio, Japan) 10-bit OLED (see Ito, Ogawa, & Sunaga, 2013 for a report on the properties of the monitor). When each channel of the monitor was measured at full power, the following CIE 1931 xyY coordinates were obtained: R = (0.6770, 0.3223, 44.037), G = (0.1926, 0.7277, 102.050), and B = (0.1408, 0.0501, 10.803). The distance between the eyes of the observer and the chin rest was about 0.6 m. The resolution of the screen was 1920 × 1080 pixels, which corresponds to 0.543 m × 0.306 m or 49° × 29° viewing angle. Together with this monitor, the Psychophysics Toolbox and Matlab version R2015a (The Mathworks Inc., Natick, MA, USA) were used for stimulus generation and presentation on a Dell (Dell Inc., Round Rock, Texas, USA) precision T3610 computer.

Color calibrations and measurements of the monitors' gamma curves were carried out with a Konica Minolta Spectroradiometer CS-2000 (Konica Minolta Holdings Inc., Marunouchi, Tokio, Japan). Stimuli were gamma corrected before they were displayed on the monitor.

#### 2.1.2. Stimuli

Two square patches were shown on a gray background in each trial of Experiment 1, 2, and 3 (Fig. 1b). Each side of the patches subtended 10° of visual angle in the horizontal and the vertical direction. The patches were displaced by the same amount to the left and the right of the center of the screen in the horizontal such that there was a distance of about 10° visual angle between the inner edges of the patches (cf. Switkes, 2008; the description of the setup used by Cao et al., 2014, suggests that they used stimuli of similar size; cf. also and Pitt & Winter, 1974). We used the CIE 1931 2° color matching functions for stimulus generation and all

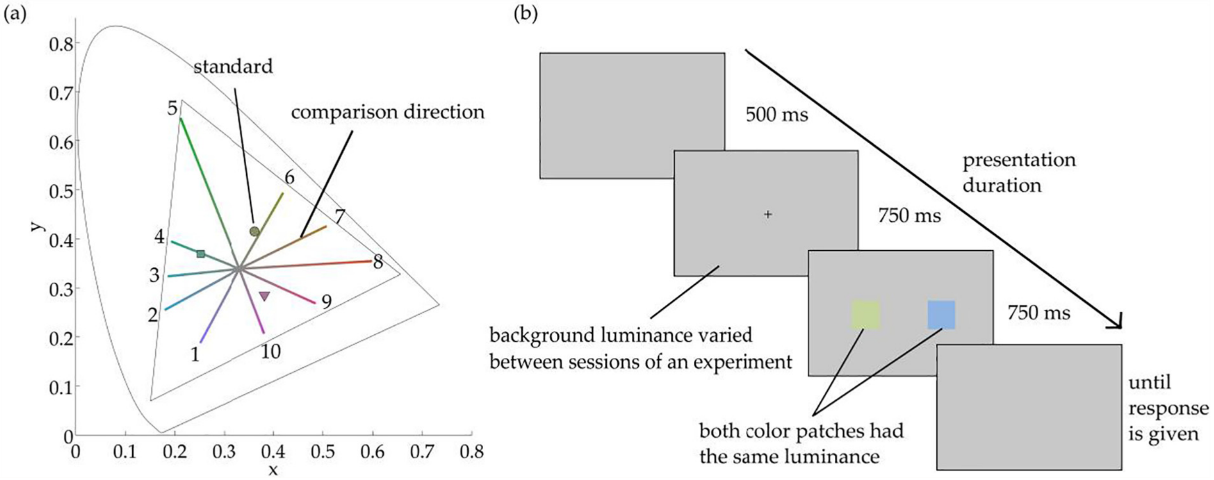

**Fig. 1.** (a) The three standard colors are represented by the greenish circle, the bluish square and the reddish triangle. Their CIE 1931 xy coordinates remained the same in Experiment 1, 2, and 3. The ten comparison directions that were used in Experiment 1 are represented by the colored straight lines. The hue angles of the directions remained the same in Experiment 1, 2, and 3. The endpoints of the directions had to be changed as luminance levels changed, however. Table 1 shows which endpoints were used in which of the three experiments. The large black triangle represents the gamut of the Eizo CG223W monitor. (b) In each trial, two color patches were shown for 750 ms. One of the patches always had the color of one of the three standards that are illustrated in Fig. 1a. The color of other patch was sampled from one of the ten comparison directions by means of an adaptive algorithm. Observers had the task to indicate which of the two patches is the more saturated by pressing one of two buttons. The luminance of the background was varied between different experimental sessions.

our computations. A reanalysis of Experiment 2 showed that the predictions of the CIE, KOE, LAB, or LUV measure did not improve by using CIE 1964 10° color matching functions. One of the patches always had the CIE 1931 xy coordinates of one of three standard colors (Fig. 1a). The coordinates of the standards were: reddish = (0.382, 0.285), bluish = (0.252, 0.369), greenish/brownish = (0.362, 0.415). The CIE 1931 xy coordinates of the other patch were sampled from one of ten comparison directions (Fig. 1a). Every color direction started in the CIE 1931 xy coordinates of the background and ended nearby the gamut of the monitor. In all three experiments, the hue angles of the color directions remained the same. Only the endpoints of the comparison directions were chosen such that the full gamut of the monitor was exploited at the levels of luminance that were used in the respective experiment. In Experiment 3, we sampled our comparison patches from only 3 of the 10 comparison directions (directions 2, 5, and 8) in order to contain the overall duration of the experiment. A list of the endpoints of the color directions in CIE 1931 xy coordinates can be found in Table 1. The gray background had the CIE 1931 coordinates x = 0.331 and y = 0.339 in all of the experiments.

The luminance of the standard and the comparison patch was always 30 cd/m$^2$ in Experiment 1, 70 cd/m$^2$ in Experiment 2, and 50 cd/m$^2$ in Experiment 3. In Experiment 1, the luminance of the gray background was 10 cd/m$^2$ in the first experimental session, and 45 cd/m$^2$ in the second. In Experiment 2, the background had a luminance of 50 cd/m$^2$ in one session and 120 cd/m$^2$ in the other. In order to keep as many variables as possible constant with regard to Experiment 1, we did not choose the monitor white as white point but continued to use x = 0.331, y = 0.339. In Experiment 3, background luminance was 30, 40, 50, 60, or 70 cd/m$^2$ in the five corresponding experimental sessions.

We used these values whenever the definition of one of the seven saturation measures required us to do so in order to determine the saturation of a patch. That is, to compute saturation of a patch with the LAB and LUV measure we used the chromaticity coordinates x = 0.331 and y = 0.339 of the gray background and the luminance of the background that pertained to a particular experimental condition (using the mean color of the whole screen improved the performance of the two measures, using the mean luminance of the screen or the maximum luminance of the monitor did not lead to a change in performance). As far as the CIECAM02 measure is concerned, we followed Luo and Li (2013) by choosing the mean luminance of the whole screen as adapting field La, by using the background luminance that pertained to a particular experimental condition as Yb, and by using the maximum luminance of the monitor RGB = (max, max, max) as reference white. These specifications are consistent with Schiller and Gegenfurtner (2016). The surround condition was specified as "average", however, since specifying it as "dim" or even "dark" did not improve performance of the measure. The coordinates x = 0.331 and y = 0.339 were also used as white point for computing saturation with the CIE and the KOE measure. Since the DKL and HSV color spaces are device dependent, their white point was given by the monitor white.

### 2.1.3. Procedure

Each trial started with a blank gray screen. After 500 ms, a fixation cross appeared on the gray background for 750 ms. The fixation cross disappeared and two color patches were then displayed simultaneously against the gray background. The patches disappeared after 750 ms and then the observer was able to respond. The screen remained gray until a response was given (Fig. 1b). The observers' task was to determine which of the two patches was the more saturated by pressing one of two buttons on the keyboard. The screen was viewed binocularly and observers could move their eyes freely during stimulus presentation.

One of the patches always had the color of one of the three standard colors, while the color of the other patch was sampled from one of the ten color directions. For each standard, each color direction, and each background illumination, an observer had to go through sixty trials, adding up to 1800 decisions in Experiment 1 and 2 (60 decisions × 10 color directions × 3 standards × 2 background illuminations) and 2700 decisions in Experiment 3 (60 decisions × 3 color directions × 3 standards × 5 background illuminations). These numbers do not include five practice trials that were presented before the actual experiment to familiarize the observer with the task. The comparison patches were sampled from the color directions by means of an adaptive algorithm in a procedure similar to the one used by QUEST (Watson & Pelli, 1983). That is, in the first twenty of the sixty trials, the color of the comparison patch was drawn at random from the corresponding color direction. Once the data of twenty trials had been collected, a cumulative Gaussian function was fitted to the data after each trial, using psignifit toolbox in Matlab (cf. Schütt, Harmeling, Macke, & Wichmann, 2015, 2016). The resulting PSE was jittered to ensure that sufficient data was sampled also from the direct surround of the PSE. The color that corresponded to this jittered PSE was the color that would be shown in the next trial. Hence, on the last forty of the sixty trials that were performed for each standard and color direction, the comparison stimulus was continuously updated in the light of the data that had been collected.

Observers received instructions in written form before the experiment. They were told that they would be shown two color patches in each trial and that they had to decide which of the two patches was more saturated. Their task was to press the left button "d" on the keyboard if they had the impression that the left patch was more saturated, while they were to press the right button "k" if they thought that the right patch is the more saturated. It was explained that "more saturated" meant as much as "more strongly colored". Observers were instructed to respond quickly and to follow their intuitive feeling about the saturation of the colors. After observers had read the instructions, they were asked to do five practice trials. After finishing the practice trials, they were asked whether they had any further questions. In case an observer was still unsure about what it meant for a color to be saturated, the following explanation was given orally: Just as you have an intuition for when a tone is louder than another, independently of its pitch, you have an intuition for when a color is more saturated than another, independently of its hue. During the experiment, we want you to resort to this intuition about saturation when you make your judgments. Observers did not receive an explicit definition of what saturation is that went beyond these clarifications.

In each trial of Experiment 1, 2 and 3, each standard color could either appear on the left or the right side. The three standards were shown in randomized order. The order in which the color of the comparison patch was sampled from the color directions was randomized as well. In Experiment 1 and 2, each of the two sessions was split into three blocks. Between every block, observers could take a break as long as they needed to be ready for the next block. All observers started Experiment 1 with the condition in which background luminance was 10 cd/m$^2$. In Experiment 2, we balanced with which background luminance our observers started and ended the experiment. That is, if observer i started the experiment with a background luminance of 50 and ended it with 120 cd/m$^2$, then observer i + 1 started it with 120 and ended it with 50 cd/m$^2$. In Experiment 3, we counterbalanced the order in which the five background luminances were used in a similar way.

### 2.1.4. Data analysis

Each participant gave 60 responses for each combination of color direction, standard color, and background luminance. The responses were grouped in 10 bins of 6 trials. The proportion of trials where a test color was judged as more saturated than the standard was computed for each bin (cf. black dots in Fig. 2b). A cumulative Gaussian function was then fitted to the data, using psignifit toolbox in Matlab (cf. Schütt et al., 2015, 2016). We used the mean of the fitted distribution as PSE and the standard deviation $\sigma$ as just noticeable difference (JND) for further computations. If a fit was "bad" in the sense that the PSE was further away from the white point than the endpoint of the direction, the PSE was set equal to the endpoint. In Experiment 1, it was necessary to do so in seven out of the overall 600 cases, five pertaining to observer 7, one pertaining to observer 2, and one to observer 8. In Experiment 2, it was necessary to do so in three out of the 600 cases, two pertaining to observer 6 and one pertaining to observer 4. In Experiment 3, in two out of the 405 possible cases the PSE had to be set equal to the endpoint (one case pertained to observer 6 and the other to observer 11). There was no case in which the fit was "bad" in the sense that the PSE went beyond the white point in the opposite direction of the endpoint.

For each of the three standards we computed the PSEs that were predicted by the seven measures for the different comparison directions and background luminances. In order to determine how much these predictions deviated from the perception of an observer, we computed the difference between the PSE and the prediction of the measures. We divided this difference by the corresponding JND:

$$d = \frac{PSE - Prediction_{measure}}{JND}.$$

We divided by the JND because we wanted the deviations of the PSE from the predictions to be comparable across different standards, color directions, participants and color spaces. In order to determine how well a measure predicted human perception of saturation on average, we aggregated the deviations d across all standards, color directions, observers, and background luminances.

Analyzing the data in this way might make the performance of a particular measure dependent on which colors are chosen as standards for the experiment. We therefore used a Nelder-Mead simplex algorithm implemented in the fminsearch function in Matlab to choose, within the boundaries that are set by the endpoints of the color directions, the level of saturation for each saturation measure that minimized the deviation from the PSEs for each standard and the two experimental conditions of Experiment 1 and 2. For instance, Fig. 2d shows the contours of equal saturation for each measure that optimally fit the PSEs that were obtained for the greenish standard when background luminance was 10 cd/m$^2$ in Experiment 1. For purposes of comparison, we ran the same analyses for these standard independent predictions that we ran for the standard-based predictions. We also used the standard independent instead of the standard-based predictions to determine in which color direction a measure performed particularly bad or well.

Furthermore, we computed a value that we call "observer consistency", which was computed for each observer and experimental condition as the distance between the observer's PSE and the average PSE of the remaining observers, again divided by the observer's corresponding JND. Formally: Let O be the set of all our observers. For each observer i we determined the mean $PSE_{O\backslash\{i\}}$ of the set $O\backslash\{i\}$ for each standard and each color direction. These mean PSEs were treated like the predictions of a saturation measure. That is, we computed the difference between the $PSE_i$ and $PSE_{O\backslash\{i\}}$ and divided this value by the JND that pertains to $PSE_i$. The resulting deviations were aggregated across all participants, color directions and standards.

To check for outliers in Experiment 1 and 2, we made use of the fact that the bluish standard is on comparison direction 4 and the greenish standard is very close to comparison direction 6. If the task was executed properly by an observer i, then the $PSE_{\{i\}}$ of observer i should not deviate much from the bluish standard on comparison direction 4 and the greenish standard near comparison direction 6. Conversely, if there is a large deviation, then this suggests that the observer clearly failed to execute the task properly. Therefore, we determined the mean $PSE_{O\backslash\{i\}}$ of the set $O\backslash\{i\}$ for the greenish standard near comparison direction 6 and the bluish standard on comparison direction 4 for each background luminance separately. We then determined the differences $PSE_{O\backslash\{i\}}$ - $PSE_{\{i\}}$. If either of these two differences was larger than three times the standard deviation pertaining to the values $PSE_{O\backslash\{i\}}$, then an observer's data was removed from data analysis. In Experiment

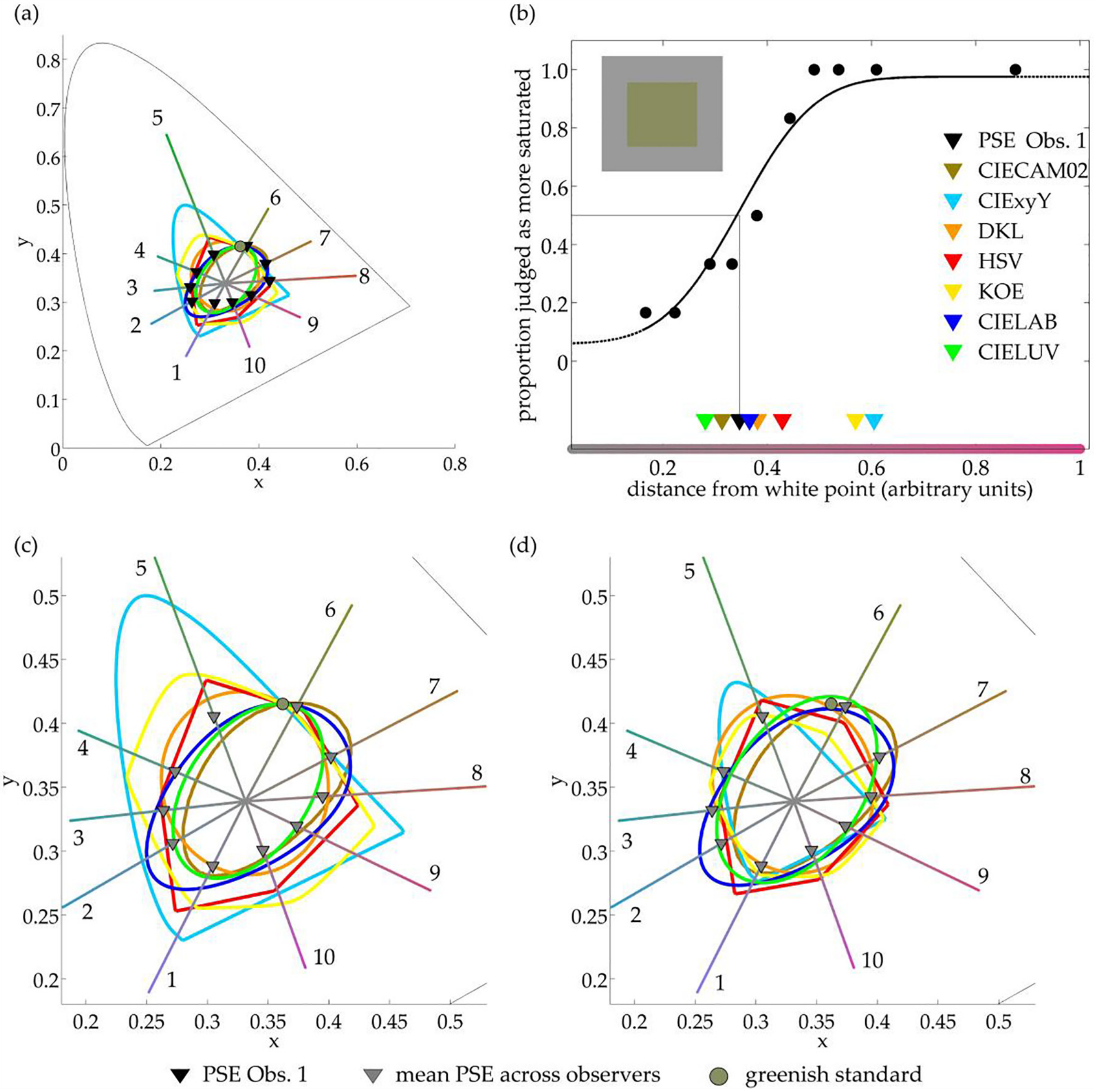

**Fig. 2.** The figures shown in (a) - (d) pertain to the condition of Experiment 1 where background luminance was lower than patch luminance (a) The numbered straight lines represent the ten comparison directions. The colored contours represent the contours of equal saturation that are predicted by the seven measures for the greenish standard, which is represented by the green dot near comparison direction 6. Note that all contours merge in the green dot because it represents the color whose saturation is predicted by the measures. The black triangles on the comparison directions represent the respective PSEs obtained from Observer 1 for the greenish standard. They are shown here for exemplary purposes. (b) The x-axis represents comparison direction 9. The black curve shows the cumulative Gaussian that was fitted to the data (black dots) that were obtained from Observer 1 for the greenish standard. The resulting PSE is indicated by the black triangle, the predictions of the measures are indicated by the colored triangles. (c) Contours of equal saturation predicted by the measures for the greenish standard. The greyish triangles represent the mean PSEs across all observers. (d) Contours of equal saturation were fitted such that their deviation from the mean PSEs that pertain to the greenish standard is minimal. The predictions of the measures thus become standard independent. Note that the contours do not have to merge in the green dot anymore.

3, outlier exclusion had to be done in a different way because line 4 and line 6 were not available anymore. We computed the mean $PSE_{O\setminus\{i\}}$ of the set $O\setminus\{i\}$ for each standard and each comparison direction and then determined the differences $PSE_{O\setminus\{i\}}$ - $PSE_{\{i\}}$. We then averaged these differences across all the comparison directions and standards for each observer. If this average difference deviated more than three standard deviations from the average difference of the other observers, then an observer's data was removed from data analysis.

The DKL and HSV measure are defined in device dependent color spaces. This means that the monitor white was used as the white point. Since the monitor white was not identical to the color that was chosen as background color in Experiment 2, however, both measures performed much worse in Experiment 2 than in Experiment 1. The insensitivity of the DKL and the HSV measure to changes in background chromaticity can be regarded as a severe shortcoming which shows their inadequacy as measures of saturation. However, since we wanted to directly compare the performance of the two measures between Experiment 2 and Experiment 1, we tried to correct for this shortcoming in our analyses. To do so, we first shifted all relevant CIE xyY values towards the monitor white of the OLED (x = 0.304, y = 0.319) by the same amount by which the chromaticity of the background (x = 0.331, y = 0.339) was off ($shift_x = 0.304–0.331 = -0.027$, $shift_y = 0.319–0.339 = -0.020$). For instance, standard three had the coordinates x = 0.362 and y = 0.415 and the endpoint of line nine had the coordinates x = 0.447 and y = 0.286. These values were shifted to x = 0.335 and y = 0.395 and x = 0.420 and y = 0.266, respectively. The PSEs on line nine were also shifted in the same way. We then performed computations for the DKL and HSV measure on these shifted values and denoted the corresponding results by DKL* and HSV*.

All statistical analyses were performed using R version 2.14.1 (R Development Core Team, 2008) and Matlab version R2012a (The Mathworks Inc., Natick, MA, USA). Greenhouse and Geisser correction (Greenhouse & Geisser, 1959) was used in order to correct for violations of the sphericity assumption whenever a factor of a repeated measures ANOVA had more than two levels. We provide the corresponding correction factor $\varepsilon_{GG}$ and corrected p-value $p_{GG}$ together with the uncorrected p-value in the following.

### 2.1.5. Participants

Thirteen observers participated in the first session of Experiment 1. Of these, only ten returned for the second session. The data of these ten observers (7 of them were female, the rest was male) met the inclusion criteria and was therefore used for data analysis. The mean age of the ten observers was 22.4 (SD = 2.6) years. Nine observers were right handed, one observer did not report on their handedness. For Experiment 2, we collected data of eleven female observers and one male observer. The data of two female observers were excluded from data analysis because they had to be classified as outliers according to our exclusion criteria. Hence, data analysis was performed on the data of the ten remaining observers. The mean age of these ten observers was 22.2 years (SD = 2.44), two of which were left-handed. For Experiment 3, we collected the data of eleven observers. Observers 2 and 10 had to be removed from analysis in accordance with the exclusion criteria. The remaining nine observers (6 of them were female, the rest male) had a mean age of 24.2 years (SD = 3.31) and all of them were right handed.

All observers had normal or corrected-to-normal vision. An Ishihara test (Ishihara, 2004) was used to very that our observers' color vision was normal. Our experiments were in agreement with the Helsinki declaration, approved by the local ethics committee (LEK 2009-0008) and all observers provided informed consent.

## 2.2. Experiment 1: Saturation in different directions of color space

### 2.2.1. Results

Fig. 3a and b show the PSEs that were obtained from Experiment 1 in the CIE 1931 diagram for when background luminance was 10 cd/m$^2$ and 45 cd/m$^2$, respectively. The large reddish triangle represents the reddish standard. The small triangles represent the colors that were judged as equally saturated as the reddish standard. The lines that connect these triangles can be interpreted as contours of equal saturation for the reddish standard. In an analogous way, the PSEs obtained for the bluish standard and the greenish standard are represented by small squares and small discs, respectively. When background luminance was 10 cd/m$^2$, then the PSEs obtained for the reddish standard were further away from the white point in each color direction than the PSEs obtained for the bluish standard. This indicates that the reddish standard was perceived as more saturated than the bluish standard on average. This relationship reversed when background luminance was 45 cd/m$^2$. Then, the PSEs obtained for the bluish standard were further away from the white point than the PSEs obtained for the reddish standard, which indicates that the bluish standard was perceived as more saturated than the reddish standard.

This reversal can also be observed on the level of individual observers. That is, the PSEs for the red standard were larger than or about as large as the PSEs for the blue standard in each color direction for seven out of ten participants when background luminance was 10 cd/m$^2$, while the reverse was true when background luminance was 45 cd/m$^2$. In order to test this reversal statistically, we determined for each observer and color direction the difference of the PSEs that pertain to the reddish and the bluish standard for background luminances of 10 and 45 cd/m$^2$. These differences were divided by the mean of the JNDs that pertain to the PSEs:

$$d_{PSEs} = \frac{PSE_{reddish\ standard} - PSE_{bluish\ standard}}{\frac{1}{2}(JND_{PSE_{reddish\ standard}} + JND_{PSE_{bluish\ standard}})}$$

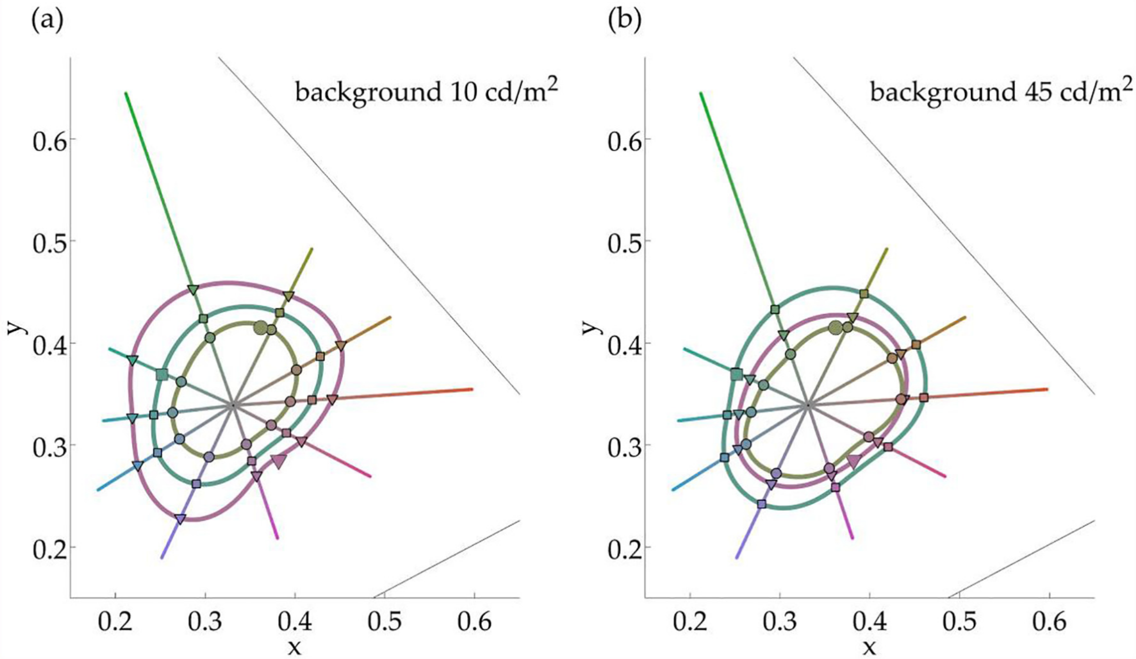

**Fig. 3.** The large reddish triangle, large bluish square and large greenish disc describe the position of the reddish, bluish, and greenish standard in CIE 1931 space, respectively. Small triangles, squares, and discs on the comparison directions represent the mean PSEs for the reddish, bluish, and greenish standard, respectively. Lines connecting the symbols are interpolations that describe contours of equal saturation. (a) PSEs for a background luminance of 10 cd/m$^2$ and (b) for a background luminance of 45 cd/m$^2$. Note that the contour of equal saturation for the bluish standard is enclosed by the contour of equal saturation for the reddish standard in (a), while this relationship is reversed in (b).

Finally, we aggregated the data across the different color directions for each observer and computed a paired *t*-test. When background luminance was 10 cd/m$^2$, then $d_{PSEs\ 10cd/m^2} = 1.57$ was significantly larger ($t(9) = 13.42$, $p < 0.001$) than $d_{PSEs\ 45cd/m^2} = -0.97$ when background luminance was 45 cd/m$^2$. Both values were significantly different from zero ($t(9) = 7.54$, $p < 0.001$ and $t(9) = -4.27$, $p = 0.002$). Hence, there was significant reversal.

This reversal is not predicted by any of the seven measures. Therefore, we determined for the two experimental conditions separately by how much the measures' predictions deviate from the judgments of our observers. Fig. 4a and b illustrate these deviations for background luminance of 10 and 45 cd/m$^2$, respectively. Fig. 4c shows the deviations averaged across the two background luminances.

A repeated-measures ANOVA with the factors "saturation measure" (levels: CIE, CAM, DKL, HSV, KOE, LAB, LUV) and "background" (levels: 10 cd/m$^2$, 45 cd/m$^2$) and the dependent variable "deviation from PSE" revealed that there was a main effect of "saturation measure" ($F(6,54) = 63.77$, $p < 0.001$, $\varepsilon_{GG} = 0.36$, $p_{GG} < 0.001$). This indicates that the measures differed in how

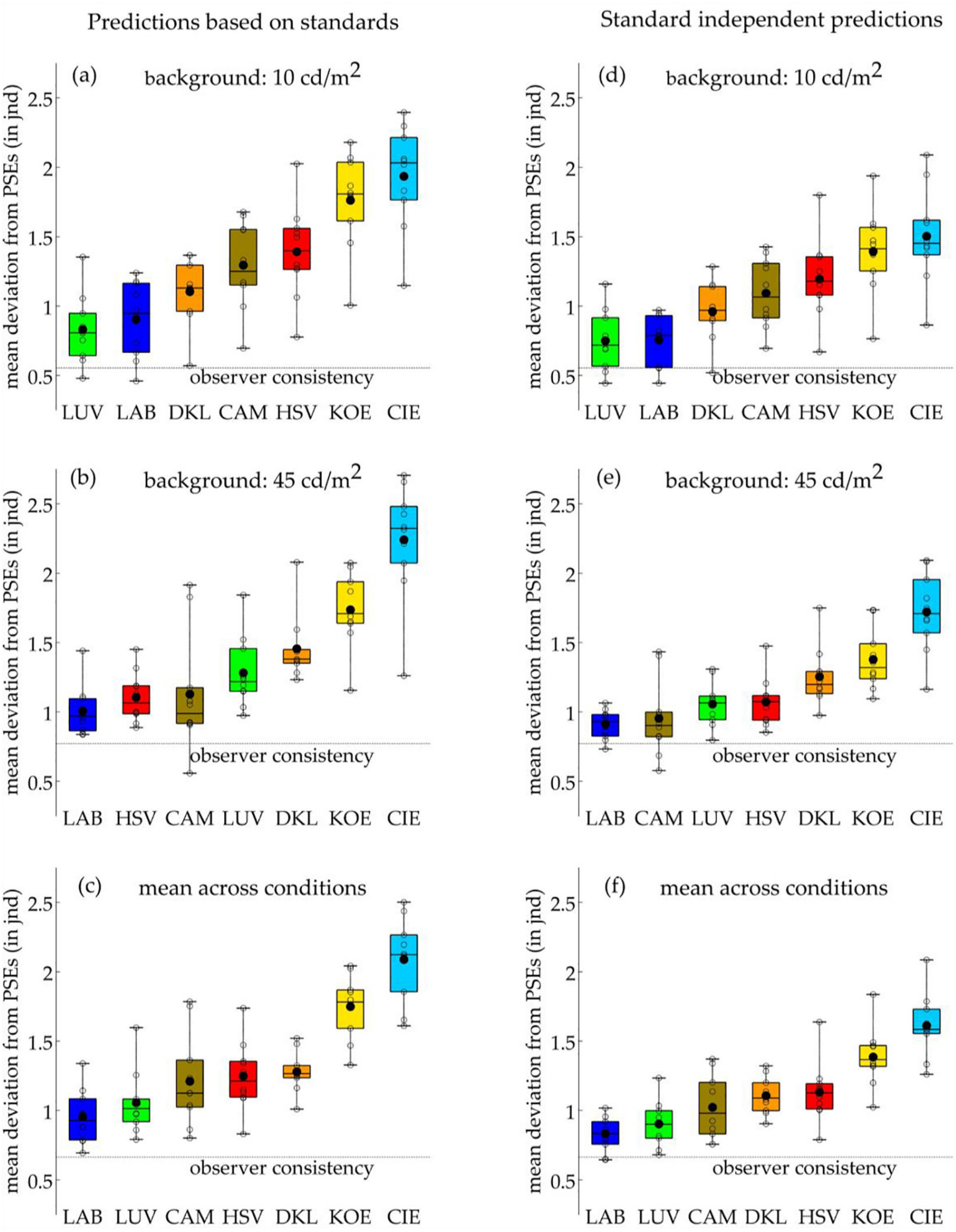

**Fig. 4.** Deviations of the predictions of the measures from the PSEs, averaged across the three standards and ten comparison directions. Small empty dots represent the means of the individual observers, the big black filled dots represent the grand mean across all observers. The central horizontal lines in the boxes indicate the median of the individual means, the whiskers indicate their range. The lower and upper end of the boxes describe the 25% and 75% percentile. (a) Deviations of the predictions of the measures from the PSEs when background luminance is 10 cd/m$^2$. (b) Deviations when background luminance is 45 cd/m$^2$. (c) Deviations averaged across conditions. Fig. 4 (d), (e), and (f) show the corresponding deviations of the standard independent predictions of the measures from the PSEs.

much they deviated from the judgments of our observers (Fig. 4). However, the main effect of "background" was not significant (F(1, 9) = 1.44, p = 0.264). The interaction between the factors "saturation measure" and "background" was significant, however (F(6, 54) = 14.90, $p < 0.001$, $\varepsilon_{GG} = 0.35$, $p_{GG} < 0.001$). This indicates that different measures performed differently for the different backgrounds (Fig. 4a and b). To be more specific, the performance of the LUV and the DKL measure became worse relative to the other measures as background luminance was increased, while the performance of the LAB, CIECAM, and the HSV measure improved.

Multiple paired *t*-tests with Bonferroni correction show that 13 out of 21 pairs of measures differed significantly from each other (Table 2). Furthermore, all of the measures performed significantly above observer consistency according to multiple paired *t*-tests with Bonferroni correction.

Conducting the same analyses on the standard independent predictions of the measures yields similar results (Fig. 4d-f). Again, there was a main effect of the factor "saturation measure" (F(6,54) = 54.61, $p < 0.001$, $\varepsilon_{GG} = 0.38$, $p_{GG} < 0.001$) and a significant interaction between "saturation measure" and "background" (F(6,54) = 10.86, $p < 0.001$, $\varepsilon_{GG} = 0.42$, $p_{GG} < 0.001$). There also was no main effect of "background" (F(1,9) = 1.76, p = 0.217). Multiple *t*-tests with Bonferroni correction showed that 15 out of 21 pairs of measures differed from each other (Table 3).

Fig. 5 shows the deviations of the standard independent predictions of the measures from the PSEs for each color direction separately. The LAB measure, which performed well on average, showed relatively high deviations in the bluish comparison direction 2, while its deviations were relatively low in the reddish directions 7 and 8. The opposite was the case for the LUV measure. The LUV measure performed relatively well in the bluish direction, while it exhibited weaknesses in the reddish direction. The CIE, KOE, and HSV measure were bad at predicting observers' saturation judgments in the greenish direction 5 and in the reddish direction 7.

In the experimental condition where background luminance was lower than patch luminance, the average illumination of the screen was only 11.5 cd/m$^2$. Although photopic vision can be expected to start above 3 cd/m$^2$ (Bayer, Paulun, Weiss, & Gegenfurtner, 2015; Stockman & Sharpe, 2006; Zele & Cao, 2015), we cannot exclude the possibility that the reversal that we found was due to the influence of rod vision. That is, when background luminance was lower than patch luminance, rod vision may still have had an influence on how the reddish and bluish standard were perceived while it is unlikely that this was the case when background luminance was higher than patch luminance (average luminance of the screen was 43.9 cd/m$^2$ in this condition). In order to exclude this possibility and to verify that the reversal is mainly due to the luminance differences between the background and the patches, we conducted a control experiment.

This experiment was the same as the earlier experiment with the only difference that we held background luminance constant at 30 cd/m$^2$ while we varied the luminance of the patches instead. That is, the luminance of both patches was 45 cd/m$^2$ in one condition, while it was 10 cd/m$^2$ in the other. We invited the seven observers who showed the reversal back into our laboratory. Three of them returned.

All observers showed a clear reversal in the sense that the reddish standard was judged as more saturated than the bluish standard when patch luminance was 45 cd/m$^2$ (Fig. 6a), while they judged the bluish standard as more saturated than the reddish standard when patch luminance was 10 cd/m$^2$ (Fig. 6b).

### 2.2.2. Discussion

In Experiment 1, we measured points of equal saturation in ten color directions while background luminance was varied and found an influence of background luminance on perceived saturation. When background luminance was 10 cd/m$^2$, the reddish standard was perceived as more saturated than the bluish standard. This relationship reversed when the brightness of the patches was decreased by raising background luminance to 45 cd/m$^2$. Then, the bluish standard was judged as more saturated than the reddish standard. These findings are coherent with data reported in Bimler et al. (2006, Fig. 1, bottom right) which suggest that a less luminant background desaturates bluish spectral lights more than reddish spectral lights, while the reverse seems to be true for a more luminant background. Our control condition where patch luminance instead of background luminance was varied indicates that the reversal is due to luminance contrast (which would also be coherent with Paramei, Bimler, & Cavonius, 1998) and can hardly be attributed to the influence of rod vision.

The reversal is not predicted by any of the measures. Thus, it is not surprising that performance of the measures varied depending on the level of background luminance. When background luminance was lower than the luminance of the patches, then the LAB, the LUV, and the DKL measure predicted the judgments of our observers best. When background luminance was higher than the luminance of the patches, then the LAB, the HSV and the CAM were the three best measures. For both background luminances,

**Table 2**
Multiple paired *t*-tests with Bonferroni correction (predictions based on the standards).

| | CAM | CIE | DKL | HSV | KOE | LAB |
|---|---|---|---|---|---|---|
| CIE | <0.001 | | | | | |
| DKL | n.s. | <0.001 | | | | |
| HSV | n.s. | <0.001 | n.s. | | | |
| KOE | 0.011 | <0.001 | <0.001 | <0.001 | | |
| LAB | n.s. | <0.001 | <0.004 | <0.001 | <0.001 | |
| LUV | n.s. | <0.001 | n.s. | n.s. | <0.001 | n.s. |

**Table 3**
Multiple paired *t*-tests with Bonferroni correction (standard independent predictions).

| | CAM | CIE | DKL | HSV | KOE | LAB |
|---|---|---|---|---|---|---|
| CIE | <0.002 | | | | | |
| DKL | n.s. | <0.001 | | | | |
| HSV | n.s. | <0.001 | n.s. | | | |
| KOE | 0.019 | <0.001 | <0.002 | <0.001 | | |
| LAB | n.s. | <0.001 | <0.002 | <0.003 | <0.001 | |
| LUV | n.s. | <0.001 | 0.011 | 0.009 | <0.001 | n.s. |

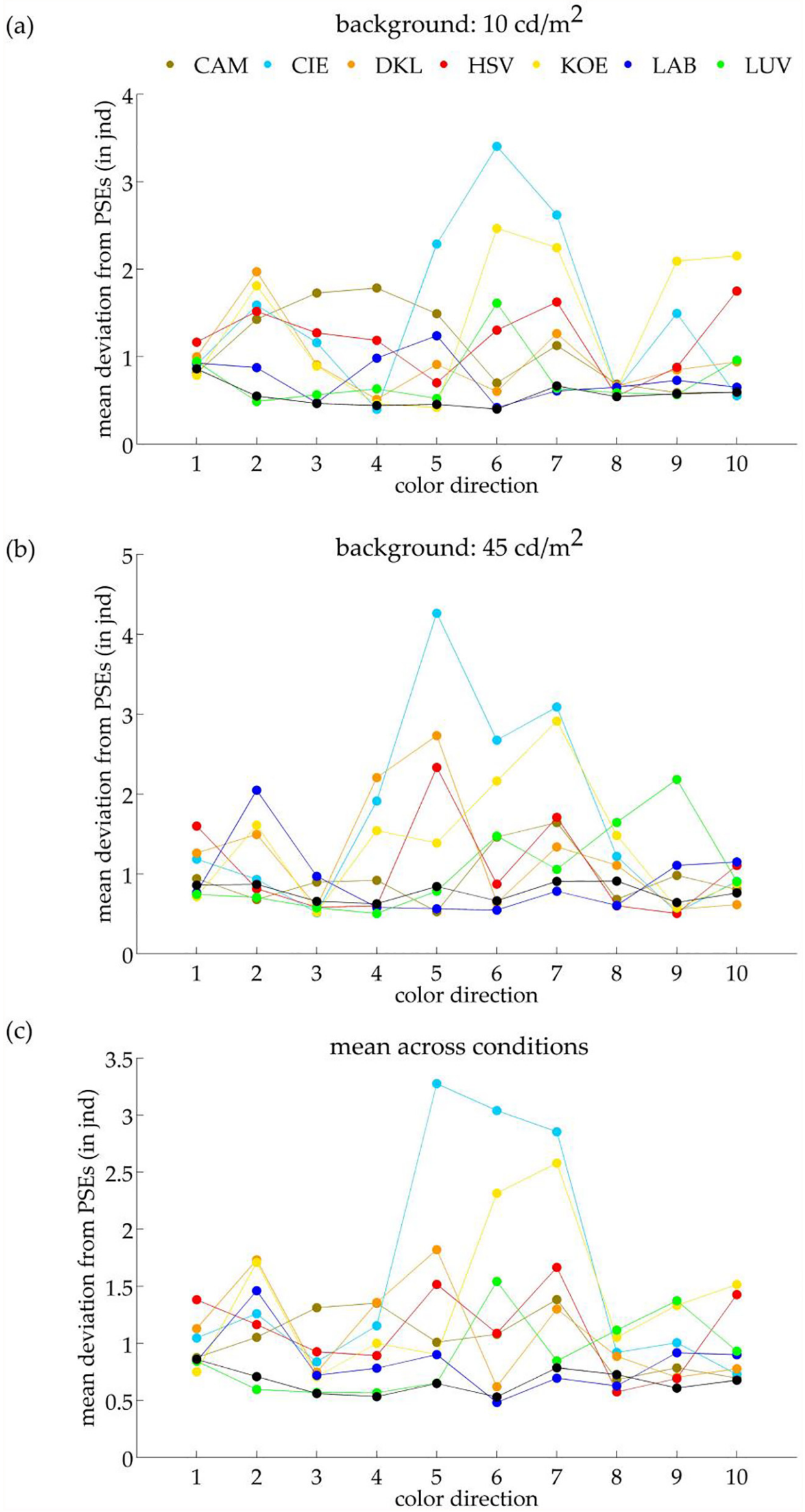

**Fig. 5.** Deviations of the standard independent predictions of the measures from the PSEs, illustrated for each color direction separately. (a) shows the deviations for a background luminance of 10 cd/m$^2$, (b) shows them for a background luminance of 45 cd/m$^2$, and (c) shows the average across conditions. Note that the deviations may be slightly below observer consistency for some measures due to the way data was aggregated.

the CIE measure performed worst, followed by the KOE measure. If one considers the average performance of the measures across the two experimental conditions, the LAB, the LUV, and the CAM measures performed best, while the KOE and the CIE measure performed worst again. Note that while the performance of all measures could be improved by making their predictions independent from three standards, this hardly changed their performance relative to each other.

### 2.3. Experiment 2: Saturation at higher luminances

#### 2.3.1. Results

The mean of the PSEs that were obtained for a background luminance of 50 cd/m$^2$ and 120 cd/m$^2$ can be seen in Fig. 7a and b, respectively. As in Experiment 1, we found that the saturation of the reddish standard was judged as higher than the saturation of the bluish standard when background luminance was lower than patch luminance, while this relationship reversed when background luminance was higher than patch luminance. Again, 7 out of 10 observers showed the reversal. We quantified the reversal in the same way as in Experiment 1. When background luminance was 50 cd/m$^2$, then $d_{\text{PSEs } 50\text{cd/m}^2} = 1.44$. This value was significantly different from zero ($t(9) = 8.40$, $p < 0.001$) and larger ($t(9) = 9.20$, $p < 0.001$) than $d_{\text{PSEs } 120\text{cd/m}^2} = -0.64$, which also differed significantly from zero ($t(9) = -3.66$, $p = 0.005$). Hence, there

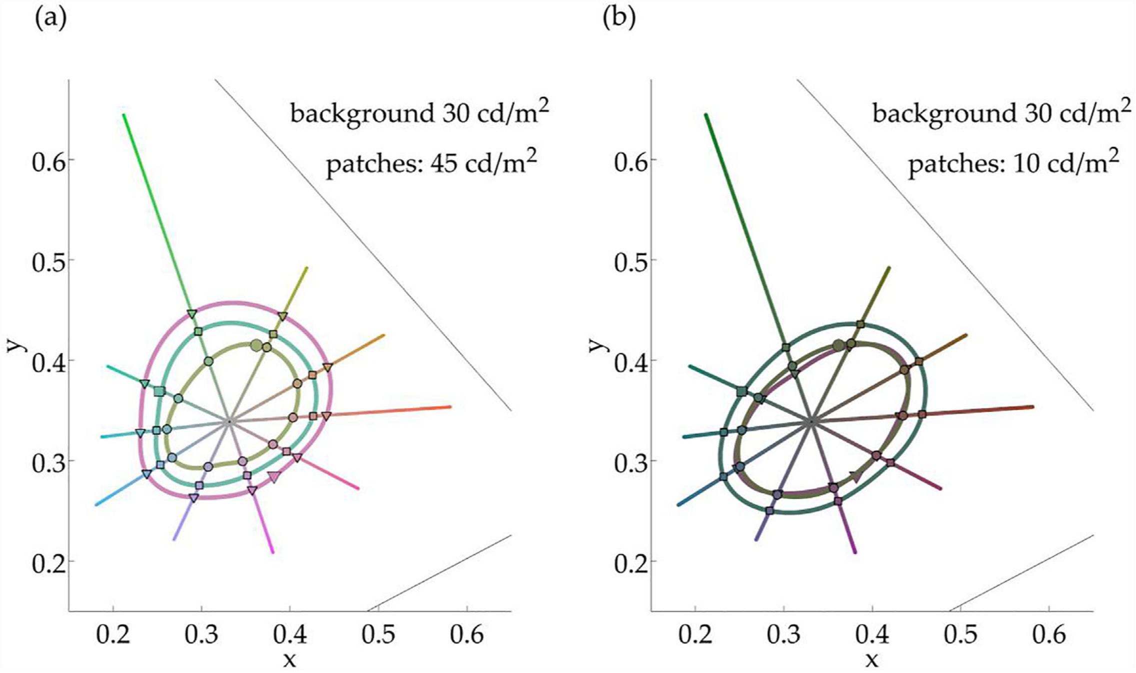

**Fig. 6.** The symbols have the same meaning as in Fig. 3. (a) shows PSEs and corresponding contours of equal saturation of n = 3 observers for a patch luminance of 45 cd/m$^2$ while (b) shows them for a patch luminance of 10 cd/m$^2$. Background luminance remained constant at 30 cd/m$^2$. Note that the contour of equal saturation for the bluish standard is enclosed by the contour of equal saturation for the reddish standard in (a), while this relationship is reversed in (b).

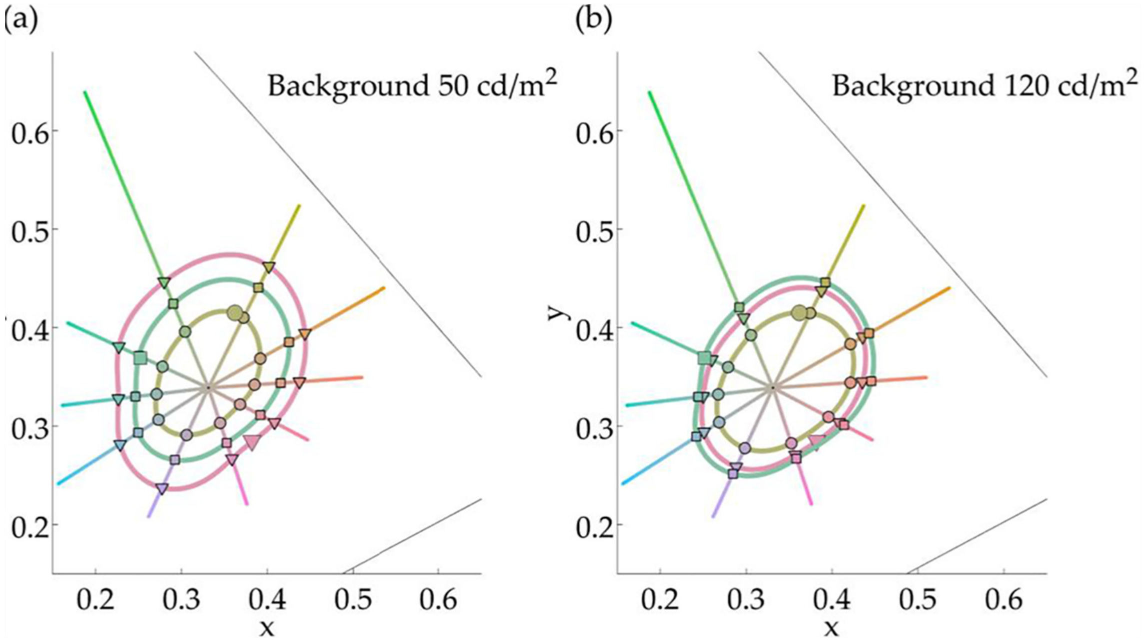

**Fig. 7.** Symbols were chosen as in Fig. 3. (a) shows PSEs for a background luminance of 50 cd/m$^2$, (b) for a background luminance of 120 cd/m$^2$. Note that the contour of equal saturation for the bluish standard is enclosed by the contour of equal saturation for the reddish standard in (a), while this relationship is reversed in (b).

was a significant reversal. To determine whether this reversal was different from the one that was obtained from Experiment 1, we ran a mixed ANOVA with the dependent variable $d_{\mathrm{PSEs}}$, the within subjects factor "background" (levels: background luminance lower than patch luminance, background luminance higher than patch luminance) and the between subjects factor "experiment" (levels: Experiment 1, Experiment 2). The interaction between both factors was not significant ($F(1,18) = 2.498$, $p = 0.131$). This suggests that the reversal in Experiment 1 was of about the same size as in Experiment 2. The main effect of "experiment" was not significant, either ($F(1,18) = 0.200$, $p = 0.660$). However, the main effect of "background" was significant ($F(1,18) = 245.621$, $p < 0.001$), as could be expected from the earlier analyses.

To test the performance of the measures, we conducted an ANOVA with the factors "saturation measure" (levels: CIE, CAM, DKL*, HSV*, KOE, LAB, LUV) and "background" (levels: 50 cd/m$^2$, 120 cd/m$^2$) and the dependent variable "deviation from PSE". This ANOVA revealed that there was a main effect of "saturation measure" ($F(6,54) = 54.70$, $p < 0.001$, $\varepsilon_{GG} = 0.38$, $p_{GG} < 0.001$). Hence, the measures differed in how much they deviated from the judgments of our observers (Fig. 8). The main effect of background was not significant ($F(1, 9) = 0.001$, $p = 0.970$). Finally, the interaction between the factors "saturation measure" and "background" was significant ($F(6, 54) = 14.17$, $p < 0.001$, $\varepsilon_{GG} = 0.31$, $p_{GG} < 0.001$). Hence, different backgrounds had a different impact on how well the measures performed (Fig. 8a and b). The performance of the LUV and the DKL* measure became worse relative to the other measures as background luminance was increased, while the performance of the LAB, CIECAM, and the HSV* improved. This was already observed in Experiment 1. Multiple

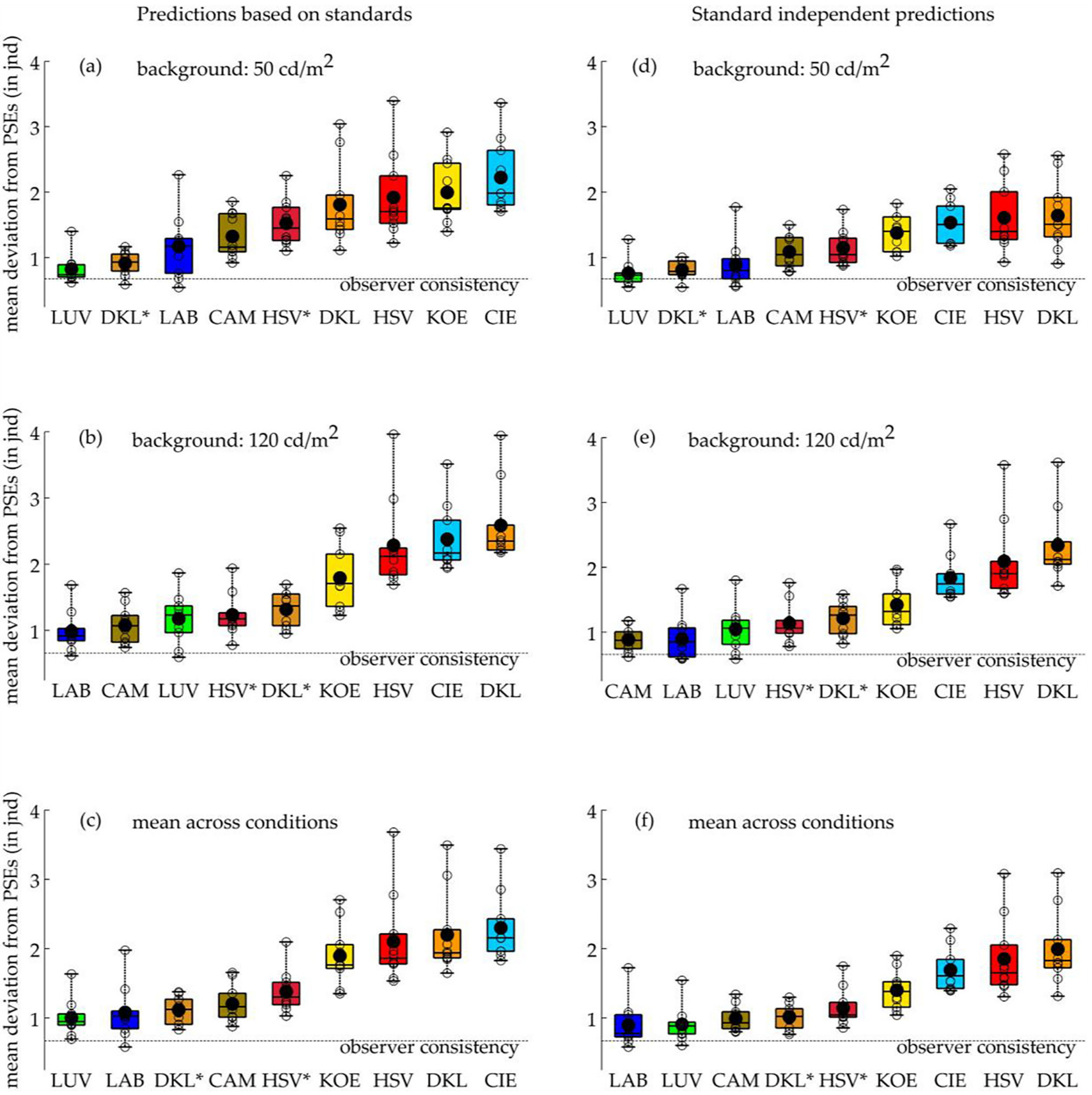

**Fig. 8.** Deviations of the predictions of the measures from the PSEs, averaged across the three standards and ten comparison directions. Symbols were chosen as in Fig. 4. (a) Deviations of the predictions of the measures from the PSEs when background luminance is 50 cd/m$^2$. (b) Deviations when background luminance is 120 cd/m$^2$. (c) Deviations averaged across conditions. Fig. 8d-f show the corresponding deviations of the standard independent predictions of the measures from the PSEs.

paired *t*-tests with Bonferroni correction showed that 13 out of 21 pairs of measures differed significantly from each other (Table 4), and that all of the measures performed significantly above observer consistency.

We also ran an ANOVA on the deviations that resulted from the standard independent predictions of the measures. The main effect of "saturation measure" was significant ($F(6,54) = 41.68$, $p < 0.001$, $\varepsilon_{GG} = 0.49$, $p_{GG} < 0.001$). Hence, the measures differed in how much they deviated from the judgments of our observers (Fig. 8d-f). The main effect of background was significant ($F(1, 9) = 7.09$, $p = 0.026$). Finally, the interaction between the factors "saturation measure" and "background" was also significant ($F(6, 54) = 16.06$, $p < 0.001$, $\varepsilon_{GG} = 0.43$, $p_{GG} < 0.001$). Hence, different backgrounds had a different impact on how well the measures performed relative to each other (Fig. 8d and e). As before, the performance of the LUV and the DKL* measure became worse relative to the other measures as background luminance was increased, while the performance of the LAB, CIECAM, and the HSV improved. Multiple paired *t*-tests with Bonferroni correction show that 13 out of 21 pairs of measures differed significantly from each other (Table 5).

Fig. 9 shows a similar pattern as was found for Experiment 1 for the standard independent predictions of the measures: In comparison to the other measures, the LAB measure performed well in the reddish direction and badly in the bluish direction, while the reverse was true for the LUV measure. The CIE, KOE, and HSV measure were bad at predicting observers' saturation judgments in the greenish direction 5 and in the reddish direction 7.

In order to compare the data obtained from Experiment 1 and Experiment 2, we conducted an ANOVA with the within subjects factors "saturation measure" (levels: CIE, CAM, DKL/DKL*, HSV/HSV*, KOE, LAB, LUV) and "background" (levels: background luminance lower than patch luminance, background luminance higher than patch luminance), the between subjects factor "experiment" (levels: Experiment 1, Experiment 2) and the dependent variable "deviation from PSE" (these are the deviations from the standard dependent predictions).

In agreement with the earlier analyses, we found a significant main effect for "saturation measure" ($F(6,108) = 113.50$, $p < 0.001$, $\varepsilon_{GG} = 0.40$, $p_{GG} < 0.001$) and a significant interaction between "saturation measure" and "background" ($F(6, 108) = 27.87$, $p < 0.001$, $\varepsilon_{GG} = 0.41$, $p_{GG} < 0.001$). All other main effects and interactions were not significant, except the interaction between "experiment" and "measure", which failed to reach significance after Greenhous-Geisser correction: $F(6, 108) = 2.60$, $p < 0.022$, $\varepsilon_{GG} = 0.40$, $p_{GG} = 0.076$. This suggests that the performance of the measures was about the same for the different overall luminance levels that were used in the experiments.

We also wanted to determine if the PSEs that resulted from Experiment 1 differed from those that resulted from Experiment 2 in their absolute distance to the white point. To do so, we conducted ANOVAs with the within subjects factor "comparison direction" (levels: comparison direction 1, 2, 3, ..., 10) and the between subjects factor "experiment" (levels: Experiment 1, Experiment 2) separately for each of the three standards and each background luminance (which could either be lower or higher than the patch luminance). This led to six ANOVAs whose results are summarized in the following. For the standards being presented on low background luminances, the main effect of "experiment" was not significant (it was closest to being significant for standard 3 with $F(1,18) = 2.27$ and $p = 0.150$), while the main effect of "color direction" always reached significance (it was closest to being not significant for standard 3 with $F(9,162) = 33.37$, $p < 0.001$, $\varepsilon_{GG} = 0.38$, $p_{GG} < 0.001$). The interaction between "experiment" and "comparison direction" was not significant for either of the three standards after application of Greenhouse-Geisser correction (it was closest to being significant for standard 1 with $F(9,162) = 2.50$, $p = 0.010$, $\varepsilon_{GG} = 0.34$, $p_{GG} < 0.067$). For high background luminances, the main effect of "experiment" failed to reach significance for each of the three standards (it was closest to being significant for standard 2 with $F(1,18) = 1.17$, $p = 0.294$), the main effect of "color direction" was significant (it was closest to being not significant for standard 3 with $F(9,162) = 32.74$, $p < 0.001$, $\varepsilon_{GG} = 0.18$, $p_{GG} < 0.001$), and the interaction was not significant (it was closest to being significant for standard 3 with $F(9,162) = 0.88$, $p < 0.544$, $\varepsilon_{GG} = 0.18$, $p_{GG} < 0.402$). Hence, raising the overall luminance level did not significantly change the absolute distance of the PSEs to the white point, the contours of equal saturation that we found in Experiment 2 were similar to those that we found in Experiment 1.

### 2.3.2. Discussion

In Experiment 2 we varied background luminance on a higher level than in Experiment 1. The contours of equal saturation were similar to those obtained from Experiment 1. A comparison of the PSEs between Experiment 1 and Experiment 2 showed that raising the overall luminance level did not have a significant effect on how distant they were from the white point. We cannot infer from this whether a Hunt effect was present or absent in our experiments.

**Table 4**
Multiple paired *t*-tests with Bonferroni correction (predictions based on the standards).

| | CAM | CIE | DKL* | HSV* | KOE | LAB |
|---|---|---|---|---|---|---|
| CIE | <0.001 | | | | | |
| DKL* | n.s. | <0.001 | | | | |
| HSV* | n.s. | <0.001 | n.s. | | | |
| KOE | 0.013 | <0.001 | <0.001 | 0.002 | | |
| LAB | n.s. | <0.001 | n.s | 0.006 | <0.001 | |
| LUV | n.s. | <0.001 | n.s | 0.002 | <0.001 | n.s. |

**Table 5**
Multiple paired *t*-tests with Bonferroni correction (standard independent predictions).

| | CAM | CIE | DKL* | HSV* | KOE | LAB |
|---|---|---|---|---|---|---|
| CIE | <0.001 | | | | | |
| DKL* | n.s. | <0.001 | | | | |
| HSV* | n.s. | <0.001 | n.s. | | | |
| KOE | 0.030 | <0.001 | 0.001 | 0.065 | | |
| LAB | n.s. | <0.001 | n.s | 0.004 | 0.005 | |
| LUV | n.s. | <0.001 | n.s | 0.017 | 0.002 | n.s. |

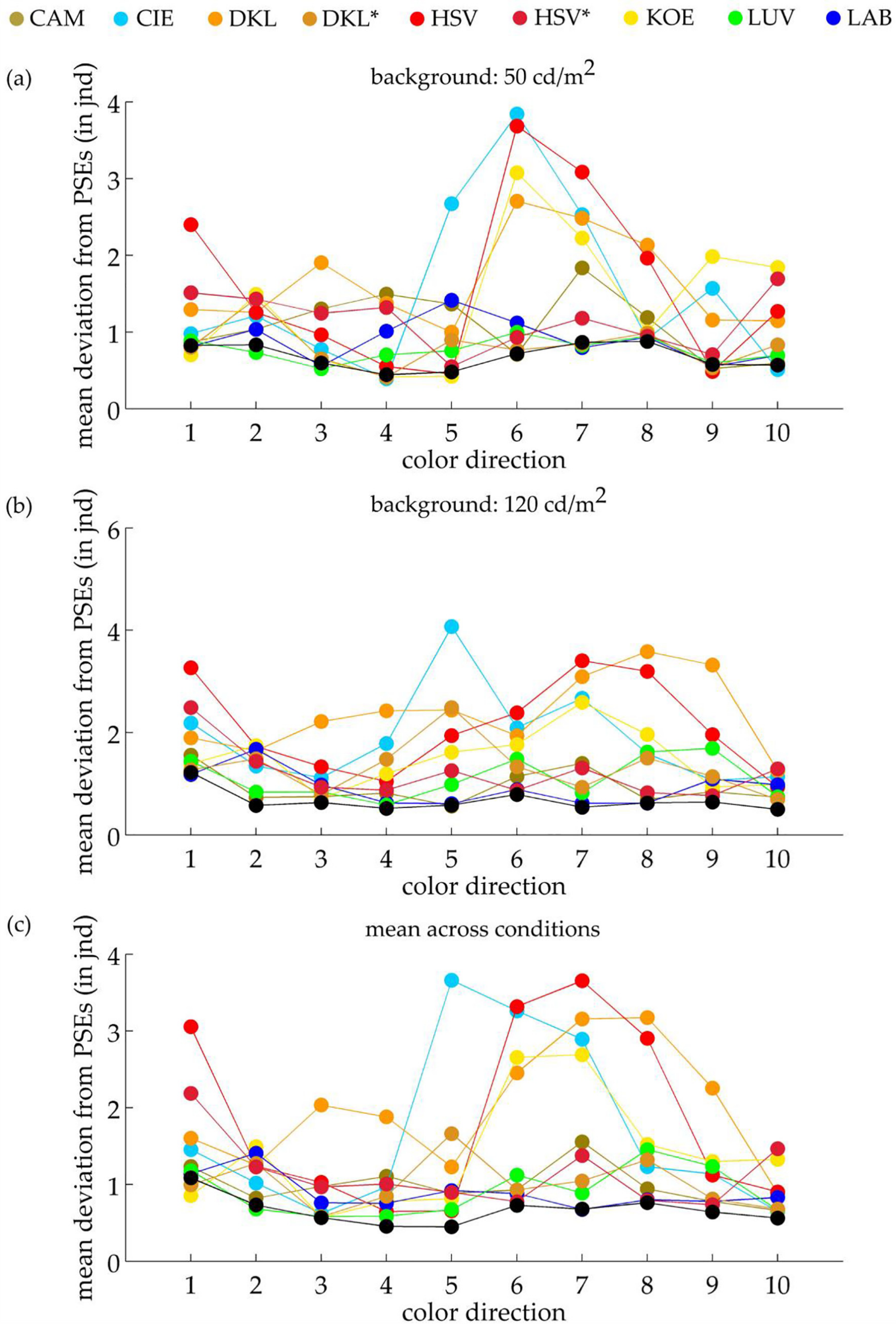

**Fig. 9.** Deviations of the standard independent predictions of the measures from the PSEs are illustrated for each color direction separately. (a) shows the deviations for a background luminance of 50 cd/m$^2$, (b) shows them for a background luminance of 120 cd/m$^2$, and (c) shows the average across conditions. Note that the deviations may be slightly below observer consistency for some measures due to the way data was aggregated.

However, we can conclude that if there was a Hunt effect, then it must have been roughly the same for all color directions.

Furthermore, we found a reversal, as in Experiment 1. That is, the reddish standard was perceived as more saturated than the bluish standard when background luminance was 50 cd/m$^2$, while this relationship reversed when the brightness of the patches was decreased by raising background luminance to 120 cd/m$^2$. The higher levels of luminance used in Experiment 2 allows us to definitively exclude the possibility that the reversal was due to rod vision.

Again, the performance of the measures varied depending on the level of background luminance. The LUV, DKL* and the LAB measure performed best when background luminance was 50 cd/m$^2$, while the LAB, CAM and the LUV performed best when background luminance was 120 cd/m$^2$. When performance of the measures was averaged across the two background luminances, then LUV, LAB, and DKL* performed best, directly followed by the CAM measure. Making the predictions of the measures independent from the three standards hardly changed their performance. The LAB, LUV, and the CAM performed best then, directly followed by the DKL* measure. Again, the CIE and the KOE measure performed worst if the uncorrected values of the HSV and the DKL measure were ignored. Since the PSEs did not differ significantly between Experiment 1 and Experiment 2, it is not surprising that the performance of the measures did not differ significantly between the two experiments either.

### 2.4. Experiment 3: Examining the reversal

#### 2.4.1. Results

For the reddish standard, the absolute distance of the PSEs from the white point decreased in all color directions as background luminance increased (Fig. 10a). For the bluish standard, the absolute distance of the PSEs from the white point decreased only for color direction 2. In the case of directions 5 and 8, the absolute distance of the PSEs rose as background luminance increased. To test whether changes in perceived saturation were brought about in a linear way, we computed the $d_{PSEs}$ values for 30, 40, 50, 60, and 70 cd/m$^2$ (Fig. 10b) and fitted a linear mixed model to these values, with background luminance as fixed effect and random intercepts and slopes for each subject. This linear mixed model appears to be well-suited for describing the decrease in dPSEs (−0.045 ±0.006) as background luminance increases (compared against a model without the fixed effect $X^2(1) = 17.21$, $p < 0.001$; $R^2 = 0.392$ for the fixed factor, $R^2 = 0.800$ for the entire model). Multiple paired *t*-tests with Bonferroni correction showed that $d_{PSEs\ 30cd/m^2}$ was significantly different from $d_{PSEs\ 50cd/m^2}$, $d_{PSEs\ 60cd/m^2}$, and $d_{PSEs\ 70cd/m^2}$. Furthermore, $d_{PSEs\ 40cd/m^2}$ was significantly different from $d_{PSEs\ 70cd/m^2}$, and $d_{PSEs\ 50cd/m^2}$ was significantly different from $d_{PSEs\ 70cd/m^2}$. However, $d_{PSEs\ 70cd/m^2} = 0.067$ was not significantly different from zero and positive, which is to say that increasing background luminance to a level that is higher than patch luminance did not lead observers to perceive the bluish standard as more saturated than the reddish standard. Hence, while there was a hue-dependent effect of changes in brightness on saturation as in Experiment 1 and in Experiment 2, there was no full reversal.

#### 2.4.2. Discussion

In Experiment 3, we varied background luminance at 30, 40, 50, 60, or 70 cd/m$^2$ while patch luminance was held constant at 50 cd/m$^2$. As in Experiment 1 and Experiment 2, we did find a hue-dependent effect of changes in brightness on saturation. That is, the difference in saturation between the bluish and reddish standard decreased as background luminance was increased relative to the luminance of the patches. This effect was brought about in an approximately linear way. Unlike in Experiment 1 and Experiment 2, however, we did not find a full reversal since the bluish standard was not perceived as more saturated than the reddish standard when background luminance was higher than the luminance of the patches. One possible explanation for this is that the contrast between the patch luminance of 30 cd/m$^2$ (70 cd/m$^2$) and the background luminance of 45 cd/m$^2$ (120 cd/m$^2$) which was present in Experiment 1 (Experiment 2) made the patches appear darker than the contrast between the patch luminance of 50 cd/m$^2$ and the background luminance of 70 cd/m$^2$ which was present in Experiment 3.

## 3. General discussion

We tested seven measures of saturation that are widely used in color science by manipulating variables known to influence perceived saturation. The measure that performed worst is the CIE measure, directly followed by the KOE measure. The predictions of the CIE and the KOE measure deviated on average by less than 2.5 JNDs from the judgments of our observers. This is surprisingly good considering that these two measures are defined in color spaces that are based on color matching functions alone and do not take the luminance of the patches or the background into account. It is less surprising that the LAB, LUV, and CAM measure performed particularly well since all three measures are defined in color spaces that are based on measurements of discrimination thresholds and take the surround into account to at least some degree. The HSV and DKL measure are both device dependent. Thus, changing the device or the chromaticity of the background strongly affects their performance. In Experiment 1, where the white point of the monitor was identical to the chromaticity of the background, the HSV and the DKL measure performed better than the KOE and the CIE measure and almost as well as the CAM measure on average. In Experiment 2, where the white point of the monitor was different from the chromaticity of the background, the HSV and DKL measure performed as badly as the KOE and the CIE measure. However, correcting for the discrepancy between the monitor white and the chromaticity of the background led to a performance of the measures (DKL* and HSV*) that was not different from the CAM measure on average.

These results are in line with the findings of Schiller and Gegenfurtner (2016) who found that the measures based on discrimination thresholds and the DKL measure are suited best for predicting saturation in natural scenes. The slight differences in performance that can be observed between our study and Schiller and Gegenfurtner (2016) can be explained well by the yellowish-bluish bias (McDermott & Webster, 2012; Nascimento, Ferreira, & Foster, 2002; Webster & Mollon, 1997) that is often found in the color distribution of natural scenes. For instance, the reason why the average performance of the LAB measure is worse than that of the CAM measure in Schiller and Gegenfurtner (2016) while the reverse is true for Experiment 1 and 2 is that it cannot predict perceived saturation in the bluish direction as well as the CAM measure (cf. Figs. 5 and 11 and Fig. 5(a) in Schiller & Gegenfurtner, 2016). Furthermore, like Schiller and Gegenfurtner (2016), we also find that the LUV measure is better than the LAB measure at predicting perceived saturation in the bluish direction.

Cao et al. (2014) found that the LUV measure performs significantly worse than the LAB and the CAM measure on average. At first sight, this finding seems to be in conflict with our results. However, one has to keep in mind that Cao et al. (2014) used Munsell patches that were always less luminant than the gray easel on which they presented their patches. Hence, the results of Cao et al. need to be compared to results that we obtained for when background luminance was higher than patch luminance. Doing so reveals the same pattern as found by Cao et al. (2014). That is, the LUV measure always performs slightly worse than the LAB

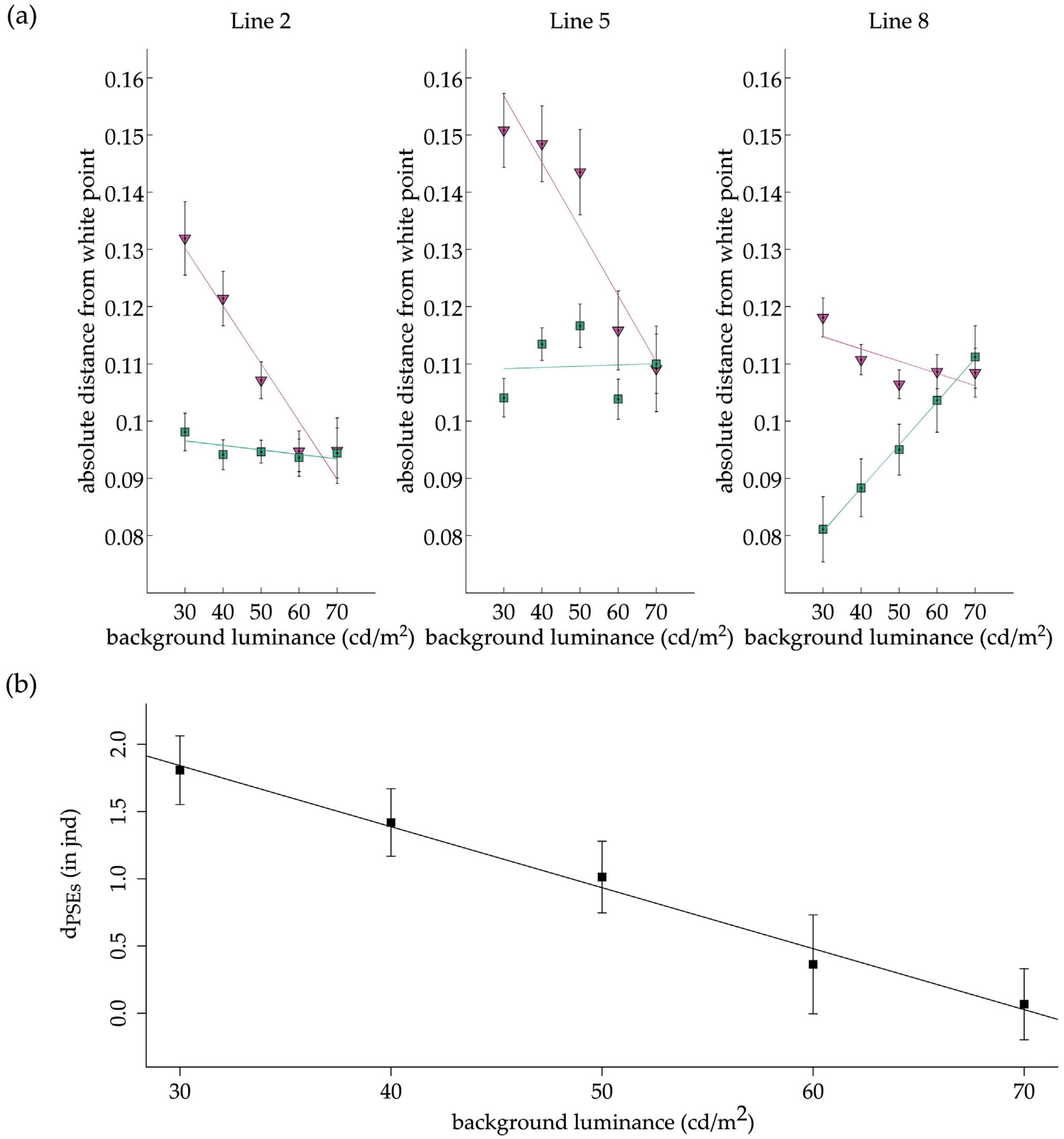

**Fig. 10.** (a) Squares represent the mean PSEs obtained for the bluish standard and triangles represent PSEs obtained for the reddish standard on comparison direction 2 (left), 5 (middle), and 8 (right). (b) Black squares represent dPSE averaged across the three comparison directions and all observers. Error bars represent one standard error of the mean in (a) and (b).

and the CAM measure when background luminance is higher than the luminance of the patches. Hence, our results agree with those obtained by Cao et al. (2014).

Note that changes in background luminance did not only affect the performance of the LUV, LAB, and CIECAM measure; the other measures were affected in their performance as well. The reason for this is that our measured contours of equal saturation changed their shape and size as background luminance was changed while the iso-saturation contours that were predicted by the measures did not (DKL, HSV, KOE and LUV) or only minimally (LAB, CIECAM). This dependence of the contours on the viewing conditions makes it difficult to compare them in detail to the results of other studies with different viewing conditions. Nevertheless, it is possible to determine whether general patterns persist across different studies. Switkes (2008, Experiment 5) found that bluish unipolar gratings required more cone contrast than yellowish gratings to be perceived as equally saturated. We found the same pattern when we represented our contours of equal saturation in DKL space. That is, they were mostly elongated along the bluish – yellowish axis (comparison directions 1 and 2 vs. 6 and 7) and shifted towards the bluish section of the S-axis. This suggests that more cone contrast was required in our experiments to make a bluish patch appear as saturated as a yellowish patch. The iso-saturation contour provided by Zemach et al. (2007) seems to be governed by a similar pattern. However, their contour exhibits a much more pronounced wedge-shape than our contours. This could be due to the different luminance contrast they used. Note that all these findings are in line with recent results showing that the bluish color direction is very special with respect to the perception of saturation (Gegenfurtner, Bloj, & Toscani, 2015; Lafer-Sousa, Hermann, & Conway, 2015; Winkler, Spillmann, Werner, & Webster, 2015; see Brainard & Hurlbert, 2015).

In our study, varying background luminance did more than just change the shape and size of the contours of equal saturation. It even affected how saturated the standards were perceived relative to each other. When background luminance was lower than the luminance of the patches, then the reddish standard was perceived as more saturated than the bluish standard. When background luminance was higher than the luminance of the patches, then the bluish standard was perceived at least as saturated as the reddish standard. None of the measures was able to predict this reversal. It is conceivable that the performance of some of the measures could be improved considerably if they were modified in order to account for such perceptual effects. However, in order to modify the measures accordingly, still more knowledge about the relationship between brightness and saturation is required.

Our study can only provide limited insights into this relationship. For instance, our results do not permit any conclusions with regard to the question whether the patches were in general perceived as more saturated when background luminance was lower than patch luminance as opposed to when it was higher. This question cannot be answered by our experiments because the saturation of the patches in one surround was not directly compared to their saturation in another. Likewise, we do not have any means of comparing perceived saturation between the three experiments, i.e. at different overall luminance levels. Thus, we can neither confirm nor disconfirm that there was a Hunt effect. However, we can conclude that *if* a Hunt effect was induced by raising the overall level of luminance, then it must have been of comparable size for all of the color directions that we examined. For, we did not find any significant differences in the absolute distances of the PSEs from the white point between Experiment 1 and Experiment 2. This is in line with Valberg (1975, p. 403) whose results suggest that the Hunt effect is the same for different color directions.

The question remains as to why we consistently found that the reddish standard is perceived as more saturated than the bluish standard when background luminance was lower than patch luminance while the bluish standard was perceived to be at least as saturated as the reddish standard when background luminance was higher. One possible answer is that this effect is an instance of what may be called a "reversed Helmholtz-Kohlrausch effect". The Helmholtz-Kohlrausch effect (Kohlrausch, 1920) consists in the fact that the brightness of a color increases as its purity is increased although its luminance stays the same (Fairchild, 1998). It is said to be a function not only of excitation purity but also of dominant wavelength (Donofrio, 2011). For instance, data provided by Wyszecki and Stiles (1982) suggest that less excitation purity is required in the direction of the reddish standard than in the direction of the bluish standard to obtain the same change in brightness. So, the Helmholtz-Kohlrausch effect seems to be stronger in the reddish than in the bluish direction. The same might be true for a reversed Helmholtz-Kohlrausch effect. Increasing the brightness of the patches would thus lead to a stronger increase in perceived saturation for the reddish than for the bluish standard so that the saturation difference between the two standards first decreases and then, from some level of brightness onwards, increases again with opposite polarity. This is precisely what we found in our experiments. Hence, it is possible to explain the reversal that we found in our experiments by a reversed Helmholtz-Kohlrausch effect.

Another explanation for the reversal can be obtained from Xing et al. (2015) who, like Faul et al. (2008) and Bimler et al. (2006, 2009), consistently found that a color becomes less saturated as the contrast between its luminance and the luminance of its surround is increased. Xing et al. (2015) measured the population response of color-sensitive neurons in V1 by recording chromatic visually evoked potentials (cVEPs) over occipital cortex while observers performed a variant of their paradigm that was suitable for EEG measurements. They found that cVEPs increased with decreasing luminance contrast. According to Xing et al. (2015), the reason for this is the following: There is a type of neurons in V1 which is responsive to luminance contrast and another type which is responsive to color and luminance (color-luminance cells). The latter type of neurons is a plausible candidate for being inhibited by the former type. The greater the luminance contrast between the patch and the surround, the higher is the inhibitory activity of the cells that are responsive to luminance contrast, which leads to decreased activity of the color-luminance cells.

The explanation given by Xing et al. (2015) can easily be extended to account for the reversal that we found. Our results suggest that the cells which respond to luminance contrast inhibit the color-luminance cells in a different way depending on the contrast's polarity. That is, when background luminance is lower than patch luminance, then the luminance cells might inhibit the color-luminance cells to a higher degree when a bluish color is shown as opposed to when a reddish color is shown. When background luminance is higher than patch luminance, then the luminance cells might inhibit the color-luminance cells to a higher degree when a reddish color is shown as opposed to when a bluish color is shown. Kinoshita and Komatsu (2001) found cells in V1 of macaque monkeys which were responsive to the contrast in luminance between a central gray patch and gray surround. Some of these cells responded in the way hypothesized by Xing et al. (2015): Their activity decreased as luminance contrast was reduced by changing the luminance of the surround. Notably, the rate at which activity changed was mostly different for a different polarity of luminance contrast (cf. Kinoshita & Komatsu, p. 2569, Fig. 10). This explanation of the reversal is in agreement with the behavioral data provided by Bimler et al. (2006, 2009) which suggest that the strength of the effect of luminance contrast on saturation depends on hue and the polarity of the contrast. Hence, it could be worthwhile to test whether our supposition is correct by using the reddish and the bluish standard in the same EEG paradigm that was used by Xing et al. (2015). If a reversal can be found in the cVEPs, then this could be evidence that the inhibition of color-luminance cells is modulated by the polarity of the luminance contrast in the way described above.

Studying cVEPs may be one important step to better understand the neurophysiological processes that determine human perception of color saturation. A complementing approach would be to test the hypothesis that saturation as a percept is the result of hue-sensitive and brightness-sensitive neurons acting together as coupled oscillators, as suggested by Billock and Tsou (2005). This hypothesis is plausible, as Billock and Tsou (2005) show, since it implies that saturation perception must follow well-established psychophysical laws. However, the electrophysiological data needed to test it are hard to acquire, as Billock and Tsou (2005, p. 2295) point out. Further work is required to explain how the percept of saturation is related to neurophysiological processes in the human brain.

## 4. Conclusion

All of the measures tested in this study are defined in color spaces that are not based on empirical measurements of saturation. Nevertheless, the predictions of the best measures (LUV, LAB, CIECAM02) deviated by only about 1 JND on average from the judgments of our observers. This suggests that the color spaces in which these measures are defined represent color saturation in a perceptually adequate way to a first approximation. The good performance of these three color spaces can be explained by the fact that they are based on measurements of discrimination thresholds and take important properties of the surround (such as its luminance or chromaticity) into account. However, there is still room for improvement. None of the measures was able to predict the hue-dependent effect of changes in luminance contrast on perceived saturation that we found in our experiments. For a successful prediction of this effect, a measure of saturation with a more complex interplay of hue and brightness is required.

## Acknowledgments

This study was funded by Deutsche Forschungsgemeinschaft (DFG) GE 879/9 and SFB TRR 135. Data pertaining to this article are publicly available at http://doi.org/10.5281/zenodo.572983. We would like to thank two anonymous reviewers for their detailed comments. Furthermore, we are grateful to David Bimler and Galina Paramei for the kind and helpful correspondence that related to their work. Finally, we would also like to thank Bianca van Kemenade and Christoph Witzel for helpful discussions.

## References

Aubert, H. (1865). *Physiologie der Netzhaut*. Breslau: Verlag von E. Morgenstern.

Bayer, F. S., Paulun, V. C., Weiss, D., & Gegenfurtner, K. R. (2015). A tetrachromatic display for the spatiotemporal control of rod and cone stimulation. *Journal of Vision, 15*(11), 15.

Billock, V. A., & Tsou, B. H. (2005). Sensory recoding via neural synchronization: integrating hue and luminance into chromatic brightness and saturation. *Journal of the Optical Society of America A, 22*(10), 2289–2298.

Bimler, D. L., Paramei, G. V., & Izmailov, C. A. (2006). A whiter shade of pale, a blacker shade of dark: Parameters of spatially induced blackness. *Visual Neuroscience, 23*, 579–582.

Bimler, D. L., Paramei, G. V., & Izmailov, C. A. (2009). Hue and saturation shifts from spatially induced blackness. *Journal of the Optical Society of America A, 26*(1), 163–172.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436.

Brainard, D. H., & Hurlbert, A. C. (2015). Color Vision: Understanding #TheDress. *Current Biology, 25*, R551–R554.

Breneman, E. J. (1977). Perceived saturation in complex stimuli viewed in light and dark surrounds. *Journal of the Optical Society of America, 67*(5), 657.

Cao, R., Castle, M., Sawatwarakul, W., Fairchild, M., Kuehni, R., & Shamey, R. (2014). Scaling perceived saturation. *Journal of the Optical Society of America A, 31*(8), 1773–1781.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology, 357*(1), 241–265.

Donofrio, R. L. (2011). Review Paper: The Helmholtz-Kohlrausch effect. *Journal of the Society for Information Display, 19*(10), 658.

Fairchild, M. (1998). *Color appearance models*. Reading: Addison Wesley Longman Inc.

Faul, F., Ekroll, V., & Wendt, G. (2008). Color appearance: The limited role of chromatic surround variance in the "gamut expansion effect". *Journal of Vision, 8*(3), 30.

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

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95–112.

Hansen, T., & Gegenfurtner, K. R. (2013). Higher-order color mechanisms: evidence from noise masking experiments in cone contrast space. *Journal of Vision, 13*, 26.

Helmholtz, H. (1852). Ueber die Theorie der zusammengesetzten Farben. *Annalen der Physik und Chemie, 163*(9), 45–66.

Hunt, R. W. G. (1950). The effects of daylight and tungsten light-adaptation on color perception. *Journal of the Optical Society of America, 40*(6), 362.

Hunt, R. W. G. (1952). Light and dark adaptation and the perception of color. *Journal of the Optical Society of America, 42*(3), 190.

Hunt, R. W. G., & Pointer, M. R. (2011). *Measuring color*. Chichester: Wiley.

Indow, T. (1978). Scaling of saturation and hue in the nonspectral region. *Perception. Psychophysics, 24*, 11–20.

Indow, T., & Stevens, S. S. (1966). Scaling of saturation and hue. *Perception & Psychophysics, 1*, 253–271.

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

Ito, H., Ogawa, M., & Sunaga, S. (2013). Evaluation of an organic light-emitting diode display for precise visual stimulation. *Journal of Vision, 13*(7), 6.

Jacobs, G. H. (1967). Saturation estimates and chromatic adaptation. *Perception & Psychophysics, 2*(7), 271–274.

Jones, L. A., & Lowry, E. M. (1926). Retinal sensibility to saturation differences. *Journal of the Optical Society of America, 13*(1), 25.

Kaiser, P. K., Comerford, J. P., & Bodinger, D. M. (1976). Saturation of spectral lights. *Journal of the Optical Society of America, 66*(8), 818.

Kim, M. H., Weyrich, T., & Kautz, J. (2009). Modeling human color perception under extended luminance levels. In *ACM SIGGRAPH 2009 papers*. New York, NY, USA: ACM, p. 27:1–27:9.

Kinoshita, M., & Komatsu, H. (2001). Neural representation of the luminance and brightness of a uniform surface in the macaque primary visual cortex. *Journal of Neurophysiology, 86*(5), 2559–2570.

Koenderink, J. J. (2010). *Color for the sciences*. Cambridge, MA: MIT Press.

Kohlrausch, A. (1920). Der Flimmerwert von Lichtmischungen. *Berichte Über Die Gesamte Physiologie Und Experimentelle Pharmakologie, 3*, 589–591.

Krauskopf, J., Williams, D. R., & Heeley, D. W. (1982). Cardinal directions of color space. *Vision Research, 22*(9), 1123–1131.

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

Luo, M. R., & Li, C. (2013). *CIECAM02 and its recent developments*. New York: Springer.

MacAdam, D. L. (1942). Visual sensitivities to color differences in daylight. *Journal of the Optical Society of America, 32*, 247–274.

Maxwell, J. C. (1857). XVIII.—Experiments on colour, as perceived by the eye, with remarks on colour-blindness. *Transactions of the Royal Society of Edinburgh, 21*(2), 275–298.

McDermott, K. C., & Webster, M. A. (2012). Uniform color spaces and natural image statistics. *Journal of the Optical Society of America, 29*, 182–187.

Nascimento, S. M. C., Ferreira, F. P., & Foster, D. H. (2002). Statistics of spatial cone-excitation ratios in natural scenes. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 19*(8), 1484–1490.

Ohta, N., & Robertson, A. R. (2005). *Colorimetry: Fundamentals and applications*. Chichester, UK: John Wiley & Sons Ltd.

Oleari, C. (2016). *Standard colorimetry: Definitions*. Algorithms and Software: John Wiley & Sons.

Paramei, G. V., Bimler, D. L., & Cavonius, C. R. (1998). Effect of luminance on color perception of protanopes. *Vision Research, 38*, 3397–3401.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Pitt, I. T., & Winter, L. M. (1974). Effect of surround on perceived saturation. *Journal of the Optical Society of America, 64*(10), 1328–1331.

R Development Core Team (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Schanda, J. (2007). *Colorimetry. Understanding the CIE system*. New Jersey: John Wiley & Sons.

Schiller, F., & Gegenfurtner, K. R. (2016). Perception of saturation in natural scenes. *Journal of the Optical Society of America A, 33*(3), A194.

Schrödinger, E. (1920). Grundlinien einer Theorie der Farbenmetrik im Tagessehen. *Annalen der Physik, 368*(21), 397–426.

Schütt, H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research, 122*, 105–123.

Schütt, H., Harmeling, S., Macke, J., & Wichmann, F. (2015). Psignifit 4: Pain-free Bayesian Inference for Psychometric Functions. *Journal of Vision, 15*(12), 474.

Stockman, A., & Sharpe, L. T. (2006). Into the twilight zone: the complexities of mesopic vision and luminous efficiency. *Ophthalmic and Physiological Optics, 26*(3), 225–239.

Switkes, E., & Crognale, M. A. (1999). Comparison of color and luminance contrast: apples versus oranges? *Vision Research, 39*, 1823–1831.

Switkes, E. (2008). Contrast salience across three-dimensional chromoluminance space. *Vision Research, 48*(17), 1812–1819.

Valberg, A. (1975). Light adaptation and the saturation of colours. *Vision Research, 15*(3), 401–404.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics, 33*(2), 113–120.

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

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*, R547–R548.

Witzel, C., & Franklin, A. (2014). Do focal colors look particularly "colorful"? *Journal of the Optical Society of America A, 31*(4), A365.

Wyszecki, G., & Stiles, W. S. (1982). *Color science: Concepts and methods, quantitative data and formulae*. New York: Wiley.

Xing, D., Ouni, A., Chen, S., Sahmoud, H., Gordon, J., & Shapley, R. (2015). Brightness-color interactions in human early visual cortex. *Journal of Neuroscience, 35*(5), 2226–2232.

Zele, A. J., & Cao, D. (2015). Vision under mesopic and scotopic illumination. *Frontiers in Psychology, 5* 1594.

Zemach, I., Chang, S., & Teller, D. Y. (2007). Infant color vision: Prediction of infants' spontaneous color preferences. *Vision Research, 47*, 1368–1381.