# Scaling perceived saturation

**R. Cao,[1] M. Castle,[2] W. Sawatwarakul,[1] M. Fairchild,[2] R. Kuehni,[1] and R. Shamey[1,*]**

[1]*Color Science and Imaging Laboratory, North Carolina State University, Raleigh, North Carolina 27695-8301, USA*
[2]*Munsell Color Science Laboratory, Rochester Institute of Technology, Rochester, New York 14623-5603, USA*
**Corresponding author: rshamey@ncsu.edu*

Received April 9, 2014; accepted June 21, 2014;
posted June 26, 2014 (Doc. ID 209771); published July 23, 2014

Two psychophysical experiments were conducted at North Carolina State University (NCSU) and Rochester Institute of Technology (RIT) to obtain replicated perceived saturation data from color normal observers on the order of one unit of saturation. The same 37 Munsell sample sheets, including up to four references that had similar perceived saturation but different hue, were used in both experiments. Different assessment methods included presenting either four references simultaneously or only one reference at a time to observers and obtaining judged saturation magnitudes for the given Munsell samples. Four saturation models comprising $S^*_{ab}$, $S^*_{uv}$, CIECAM02, as well as Richter/Lübbe, were tested. CIECAM02 gave the best prediction of saturation for data obtained at NCSU while $S^*_{ab}$ outperformed other models for the RIT data. For the combined dataset, $S^*_{ab}$, the Richter/Lübbe, and CIECAM02-based saturation models exhibited comparable performances. The Standardized Residual Sum of Squares index was used to measure the inter- and intra-observer variability and goodness of fit. Inter- and intra-observer variability of assessments was smaller than or comparable to those reported for the typical color difference evaluation experiments. 

*OCIS codes:* (330.1710) Color, measurement; (330.5020) Perception psychology; (330.1730) Colorimetry.
http://dx.doi.org/10.1364/JOSAA.31.001773

## 1. INTRODUCTION

The term "hue" is generally defined as denoting an "attribute of visual perception according to which an area (in the visual field) appears to be similar to one of the colors named red, yellow, green, or blue, or a combination of adjacent pairs of these colors, considered in a closed circle." So-called achromatic colors lack a hue and are named white, gray, or black. They differ in terms of perceived lightness. These are two of the three dimensions generally taken to define the multitude of perceived colors of objects. Hues can be considered to have a qualitative aspect, e.g., the orangeness of orange, but also a quantitative aspect, from a very weak orange to orange at its highest chromatic intensity. Viewing an orange-colored object in daylight, in natural surroundings, results in a somewhat weaker experience of orangeness. The former is considered to provide the perception on an absolute basis with the three dimensions being hue, brightness of the light viewed directly or reflected from a white surface, and what today is called colorfulness. This term was proposed by Hunt [1] in 1977 for the purpose of designating "an absolute subjective chromatic response." On a relative basis the related terms are hue, lightness, and chroma, with lightness defined as relative brightness: the brightness of an area of an object in a field of view relative to the brightness of another object that is perceived as white. Chroma is defined as the colorfulness of the object compared to that of the object appearing to be white. The term "chroma" was introduced at the beginning of the 20th century by A. H. Munsell to designate chromatic intensity of object colors at levels of equal lightness.

In 1705, in his book *Opticks*, Newton [2] used the term "intensiveness" to describe the perceptual property designated by radial lines from the achromatic center to the periphery of his hue circle. Helmholtz [3] used the term Sättigung (saturation) and, in the first edition of his *Handbuch der physiologischen Optik* in 1867, appears to have been the first to introduce the idea of a color cone with black at the endpoint. In his 1874 book for artists, W. von Bezold included a more detailed depiction of the color cone with colored illustrations of the view toward both ends. A psychophysical version of a color cone was sketched by Richter *et al.* [4] in 1940, who drew lines of constant saturation into a cross section through the Luther–Nyberg object color solid (Fig. 1).

This figure indicates that a color cone is an idealization and that the true picture is more complicated. Asked in the 1940s to develop a new kind of perceptually uniform color atlas Richter experimentally determined a constant-saturation contour in the 1931 CIE chromaticity diagram and built the system around it. The contour bears reasonable resemblance to the Munsell value 6/chroma 8 contour of the Munsell system. But the Munsell system is a cylindrical, not a conical, representation of chromatic intensity. The result of the effort is known as the DIN6164 system. The atlas was developed based on limited experimental data, intra- and extrapolation of those data, and an attempt to generate a perceptually uniform system by transforming the chromaticity data in the CIE $x$, $y$ diagram into MacAdam's 1937 $u$, $v$ version of Judd's perceptually more uniform UCS system of 1935 [5,6].

As is evident, over the years the definition of the term "saturation" has undergone a number of changes. With the introduction of the term "colorfulness" in an absolute sense and with "chroma" defined as "colorfulness as a proportion of the brightness of a similarly illuminated area that appears white" the present definition for "saturation" is "colorfulness of an area judged in proportion to its brightness," in a sense

 

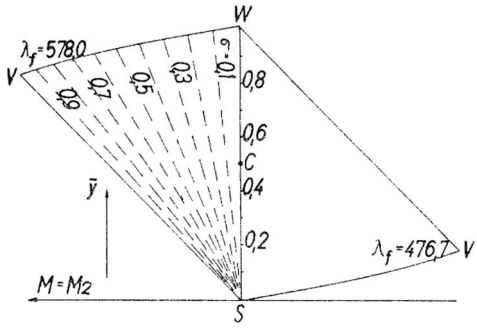


Fig. 1. Lines of constant saturation drawn into a cross section through the Luther–Nyberg object color solid [4].

comparable to that of the DIN6164 system. Expressed in terms of the Munsell or colorimetric systems, some numeric definitions of saturation are as follows:

Munsell system:

$$S = \frac{C}{V}.$$

CIELUV/CIELAB systems:

$$S^*_{uv} = 100\frac{C^*_{uv}}{L^*}, \qquad S^*_{ab} = 100\frac{C^*_{ab}}{L^*}.$$

CIECAM02 system, where $M$ is an expression for colorfulness and $Q$ represents brightness [7]:

$$S = 100\sqrt{\frac{M}{Q}}.$$

Richter/Lübbe formula: Lübbe interpreted a general concept by Richter as expressible in the following formula [8]:

$$S^+ = 100\frac{C^*_{ab}}{\sqrt{L^{*2} + C^{*2}_{ab}}}.$$

It is of some interest to compare the relationship between Munsell chroma steps for two essentially complementary hues and the CIELAB, CIELUV, and Lübbe formula results (see Fig. 2). Based on the Munsell renotation of the optimal object color limits at value 8 for hue 5Y and at value 4 for 2.5PB, essentially linear relationships are obtained between Munsell chroma and both CIELAB and CIELUV $S^*$, while for the blue color only $S^*_{ab}$ is linear. Here, the relationship between Munsell chroma and $S^*_{uv}$ is strongly nonlinear because of the compression of yellow and greenish yellow hue stimuli in the $u$, $v$ chromaticity diagram. The relationship is highly nonlinear between the Lübbe formula data and Munsell chroma for both hues. The results indicate that these three formulas predict quite different saturation values for identical stimuli.

The concept of saturation is of theoretical and practical importance, but it is perceptually not as intuitively accessible as that of chroma. When viewing, e.g., 2.5PB 2/4 and 5/12 blue samples under the same viewing conditions it is not easy to comprehend the essential identity of saturation as predicted by $S^*_{ab}$, even though blueness of the darker sample is clearly visible in good illumination. It is even more difficult to conceptually accept that samples 5Y 3/4 and 8/14 have nearly identical saturation. It is equally difficult to comprehend, when viewing the sample series 2.5PB 9/2 to 2/2, that the saturation between the lightest and the darkest sample increases by a factor of nearly 7. These seem to be abstract facts not perceptually evident at first glance.

There have been relatively few perceptual experiments that attempt to assess which formula is in best agreement with extensive experimental perceptual determinations of saturation. A relatively recent evaluation is that by Juan and Luo in 2000 [9]. For the purpose of saturation judgments, Juan's samples consisted of 132 NCS samples of different hues, lightness, and chromaticness, presented as cubes against white, light gray, and black surrounds. Seven observers were instructed in the meaning of the attributes, hue, lightness, colorfulness, and saturation and, in case of saturation, observers made comparison judgments against three different-hued reference samples. They were given a saturation value for the reference sample and judged the saturation of the test samples against it on an open scale. The observers were given extensive training in judging saturation. Mean results for all observers were compared against several appearance models available at the time, with the best fit obtained with the LLAB96 model (with the mean coefficient of variation for the three surrounds of 21%). Of all four attributes estimated in the general experiment, saturation was found to have the highest inter-observer variability. A conclusion was that observers can be trained well to make comparable saturation judgments.

Using Munsell atlas samples for an experiment assessing perceptual saturation is useful because, at a given value level, all samples have identical colorimetric lightness, differing only in chroma, unlike in the case of the NCS atlas samples. While chroma steps are not closely related to saturation, sample series offer the opportunity to determine the mean perceived saturation at a given lightness level as falling on one or between two neighboring chroma steps. A related benefit is the ability to see to what degree, if any, there is confusion by the observers about the sequence of constant value samples in regard to saturation (and implicitly chroma).

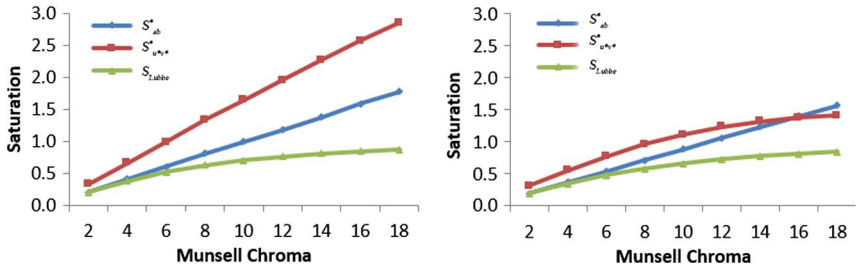


Fig. 2. Munsell chroma versus saturation for 5Y (left) and 5PB (right).

The purpose of the present experiment was twofold: (a) to establish new data of saturation judgments, based on evaluating Munsell atlas samples against samples designated, based on results from a preliminary experiment, as having an arbitrary saturation value of 1; (b) determine the replicability of results by performing the same experiment using the same methodology and samples in a different location with different experimenters and observers.

Samples of ten different hues were used in the experiment. For each hue, at a given colorimetric lightness, there were three or four samples differing in chroma around the sample with a Munsell value and chroma-related saturation value near 1. Thus, this experiment is limited to the Munsell atlas samples with the highest chroma. The methodology does not provide a complete assessment of the perceptual concept of saturation but provides data for a limited set of conditions. Assessing in a similar manner the saturation of high and low colorimetric lightness samples is, for most hues, not possible because of the limited number of samples available in the Munsell atlas at high and low values. An additional possible experiment would be to assess, for as many as 40 hues, intermediate saturation on the basis of the mean 1-unit data established in this experiment, for example at 0.5 units of perceived saturation.

## 2. METHOD

### A. Samples

A total of 37 Munsell atlas samples were obtained from X-Rite Ltd. Several $2'' \times 2''$ samples were cut from the same Munsell sheets and used in both locations for the original and replicated experiment. All samples were measured with a DataColor SF600 spectrophotometer with a large area view aperture (30 mm), and UV and specular light excluded. Samples were rotated 90° and repositioned after each reading to reduce measurement error. Each sample was measured a total of 4 times and the results were averaged. Illuminant D65 and the CIE 1964 Supplementary Standard Observer were used for all colorimetric calculations. The colorimetric attributes of the samples, $L^*a^*b^*$, are given in Table 7 in Appendix A, with the four reference samples identified in bold letters. The location of samples in the CIELAB $a^*b^*$ diagram is shown in Fig. 3. Colorimetric data of the samples were also obtained before, during, and after the experiment, with a mean change by sample of 0.28 $\Delta E^*_{ab}$ units.

### B. Sample Presentation

At NCSU samples were presented to observers in a SpectraLight QC calibrated viewing booth with a single test sample presented above the four reference samples on a light gray easel with its surface at a 45° angle relative to the booth surface and illuminated from above (as shown in Fig. 4) with a calibrated filtered tungsten approximation of D65 illumination at an intensity of 2200 lux and color temperature of 6560 K. The measured $L^*a^*b^*$ values of the easel are 72.51, −1.03, and 0.04, respectively, and those of the surrounding booth surface are 75.71, −0.41, and 1.31. A PTFE white standard was placed in the booth where samples were presented and the white point of the light source was measured using a Photo Research PR670 spectroradiometer. The absolute and relative tristimulus values of the background, easel, and PTFE are shown in Table 1. The presentation of the samples in the original experiment and its replication was essentially identical. However, the viewing booth used at RIT was a GretagMacbeth SpectraLight III set on "Daylight 75" with a CCT of 7500 K and an illuminance of 950 lux, as measured by a Konica Minolta CS-100A. The easel used was a copy of that used at NCSU with similar $L^*a^*b^*$ coordinates. While the actual light sources employed in experiments were different from the standard illuminants used for colorimetric calculations, results based on using the spectral power distribution of actual light sources are only slightly different from those based on standard illuminant data and do not alter the main findings and conclusions drawn here. As such, only the standard illuminant data are used for all calculations reported in the present work.

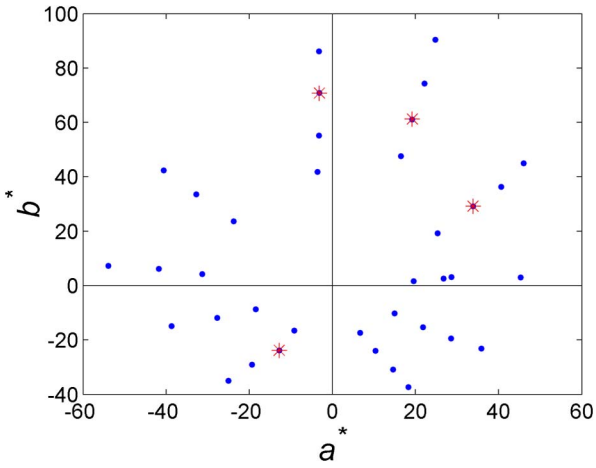


Fig. 3. Distribution of samples in the CIELAB $a^*b^*$ plane (red stars denote the location of the four references).

### C. Observers

The number of observers in the experiment at NCSU was 28 [14 male (M), 14 female (F), average age 27], while 20 observers (10 M, 10 F, average age 30) were employed in the replication at RIT. All observers were tested and found to have normal color vision. In the NCSU experiment, 72% of observers had little or no previous experience in making color related judgments, 52% were of Asian, 17% of Middle Eastern, and 31% of Western ethnicity. In the RIT replication, 45% of observers had little or no experience in color judgments and 90% of the observers were Western and 10% were Asian.

Fig. 4. Visual assessment involving 45/0 illumination viewing geometry, and a custom made sample stand painted in neutral gray that housed the standard and test samples.

**Table 1. Normalized Tristimulus Values of the Easel, Booth Surface and the PTFE Plate at NCSU**

| | $X_{10}$ | $Y_{10}$ | $Z_{10}$ |
|---|---|---|---|
| PTFE | 95.82 | 100.00 | 111.18 |
| EASEL | 47.30 | 49.41 | 53.58 |
| BOOTH | 42.22 | 44.42 | 49.35 |

The observers were requested to read an information sheet about the experiment before the first test, including: (1) illustrations of the concept of saturation via a DIN color order system on a color calibrated display [10]; (2) the procedure of the experiment and the task of the observers as described in Appendix B. Before each test observers were exposed to the experimental setup for at least 5 minutes to adapt them to the viewing conditions.

### D. Test Procedure

At NCSU, after the adaptation period, two types of assessments were carried out. First, single samples were placed in random order on the easel and the observer was asked to assess the saturation of the test sample in relation to the four references displayed below the test sample, each considered to have a saturation value of 1.0 based on results of a preliminary experiment described in this section. The results could be expressed in fractions of 1 or multiples and fractions. Following the completion of this task, references 1 and then 3 were placed below the test sample, one at a time, and the observer was asked to repeat the assessment of all samples based on each reference. Thus, each observer assessed each sample's saturation value in one trial three times. In the replication experiment at RIT, however, all four references were presented at the same time and observers gave four ratings of each test sample based on the references, assuming that the saturation value of each reference sample was 1. Each observer performed the test sequence of all samples three times, with at least 24 h between tests. If, during assessments, observers wanted to reacquaint themselves with the concept of saturation, due to having difficulty in assigning values, they were allowed to view the illustrative examples on the computer display and review the information sheet. Several observers expressed that assigning numerical values to saturation was difficult. About one third requested to view the instructions again in the second trial but felt more comfortable completing the task in the third trial.

### E. Measure of Fit

The standardized residual sum of squares (STRESS), shown in Eq. (1), was proposed by Garcia *et al.* [11] as a tool to determine the goodness of fit for the visual data and predicted data and the statistical significance of the differences between models, and to evaluate the inter- and intra-observer variability in perceptual studies [12].

$$\mathrm{STRESS} = 100\left(\frac{\sum (S_i - F_1 V_i)^2}{F_1^2 V_i^2}\right)^{1/2}, \quad \text{where } F_1 = \frac{\sum S_i^2}{\sum S_i V_i}. \tag{1}$$

Here, $V_i$ and $S_i$ are the visual and computed saturation for the $i$th sample, and $F_1$ is a scaling factor. A key property of this statistical tool is its symmetry whereby $S$ and $V$ can be interchanged. In addition, STRESS is confined to the range of 0–100, where larger values mean worse agreement between visual and computed saturation and vice versa. For a given visual dataset, the ratio of the square STRESS values from two saturation models, shown in Eq. (2), follows an $F$-distribution, and can be used to compare the statistical significance of two formulas at any confidence level:

$$F = \frac{\mathrm{STRESS}_A^2}{\mathrm{STRESS}_B^2}. \tag{2}$$

A critical $F$ value, $F_c$, which can be obtained from a lookup table or calculated, is the lower value of a two-tailed $F$ distribution with 95% confidence level, where $F_C = \mathrm{f}(\mathrm{dfA}, \mathrm{dfB}, 0.025)$, and dfA and dfB are the degrees of freedom. In this study, $F_C = 0.51$ and $1/F_C = 1.96$ [13]. Despite its clear advantages over other metrics, it should be noted that STRESS is based on a linear model that crosses the origin, and should, thus, be used with caution [14].

To determine inter-observer variability, STRESS was calculated between the mean visual ratings from repeated trials of a given observer and the mean visual ratings obtained from all observers' evaluations. For intra-observer variability, STRESS was computed for the visual ratings of each observer in each trial and the mean visual ratings from three trials for the same observer. For the performance evaluation of various saturation formulas, the overall arithmetic mean visual ratings from all observers and the predicted saturation values are used. Calculation of the results based on geometric means was not found to change the results significantly.

### F. Selection of Reference Samples

The reference samples were selected on the basis of a preliminary test at NCSU involving five color normal observers using an identical experimental procedure, except that in addition to showing all four references simultaneously in part one of the experiment (Mtd1) they repeated the test using each of the four references individually (Mtd2). In the preliminary test, 10R4/8, 10G4/8, 10GY5/8, and 10PB3/6 were employed as references. The criteria for the selection of references were: (1) selected samples should exhibit close visual ratings based on Mtd2 and (2) they should differ significantly in hue. The experimental results for the methods described showed similar intra- and inter-observer variability. In terms of STRESS inter-observer variability was 19.31 (Mtd1) and 16.61 (Mtd2) while intra-observer variability values were 17.68 (Mtd1) and 21.24 (Mtd2). The overall mean perceived saturation based on the above two methods is shown in Fig. 5. Based on the above criteria and using the evaluation results of the preliminary experiment samples 10R4/8, 10YR7/10, 10Y7/10, and 10B3/6 were selected as references showing visual saturation values of 1.365, 1.360, 1.412, and 1.343, respectively, and in the following visual experiment were assigned a saturation value of 1 unit.

## 3. RESULTS

### A. Inter- and Intra-Observer Variability

The degree of inter- and intra-observer variability reflect the "accuracy" of assessments by a given observer and "precision" among a group of observers, respectively. An "accurate observer" is one that agrees closely with mean visual results

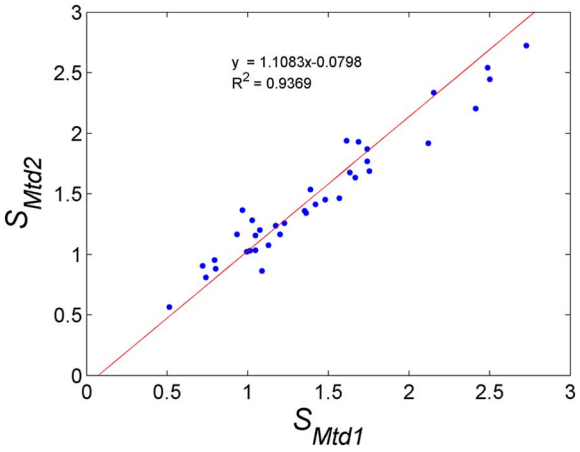


Fig. 5. Scatter plot of visual results for the two methods examined in the preliminary experiment.

from all observers, assuming that the mean values were the "true" values for each sample.

The inter-observer variability results at NCSU are shown in Fig. 6(a). They are comparable for the two different methods examined. The mean, maximum, and minimum STRESS values are 15.0, 33.0, and 6.5, respectively. Intra-observer variability results at NCSU and RIT are given in Table 2. Variability based on Mtd1, when four references were presented simultaneously, was found to be larger than that for Mtd2. Little variability in visual assessments for each observer was noted when using only one reference at a time. The inter-observer variability results of the replicated experiment at RIT are also shown in Fig. 6(b). The mean inter-observer variability was 22.0, which is 7 units larger than that at NCSU. This may be due to the fact that observers gave four ratings per sample at a given time. Also, fewer observers were employed in the RIT experiment resulting in a less accurate estimate of the mean. Intra-observer variability at RIT followed the same trend as that at NCSU. The mean STRESS value representing the intra-observer variability for four references was nearly the same (18). Intra-observer variability for the first trial was found to be larger than that for the second and the third trials. This has been reported for several psychophysical experiments previously [11,12], and has been linked to potential observer "training."

### B. Statistical Analysis of Visual Results

To analyze the precision of the mean responses and variability of the visual results, standard error (SE) and standard deviation (SD) were computed and compared. The average, maximum, minimum, and SE and SD, are listed in Table 3.

The average SD ranges from about 0.26 to 0.45 and the average SE is from 0.03 to 0.05, which is smaller than or equal to that of typical psychophysical experiments pertaining to color difference evaluation [15,16]. The results in Fig. 7 indicate that the variability increases with an increase in saturation or chroma for samples of the same hue. The variability is also hue dependent. The 10YR and 10B samples exhibit the largest mean SD (~0.45) for both methods, while 10P, 10PB, and 10Y exhibit the smallest mean SD value (~0.26) depending on the method used.

Visual saturation results of the four references were compared for the various assessment methods examined. The four

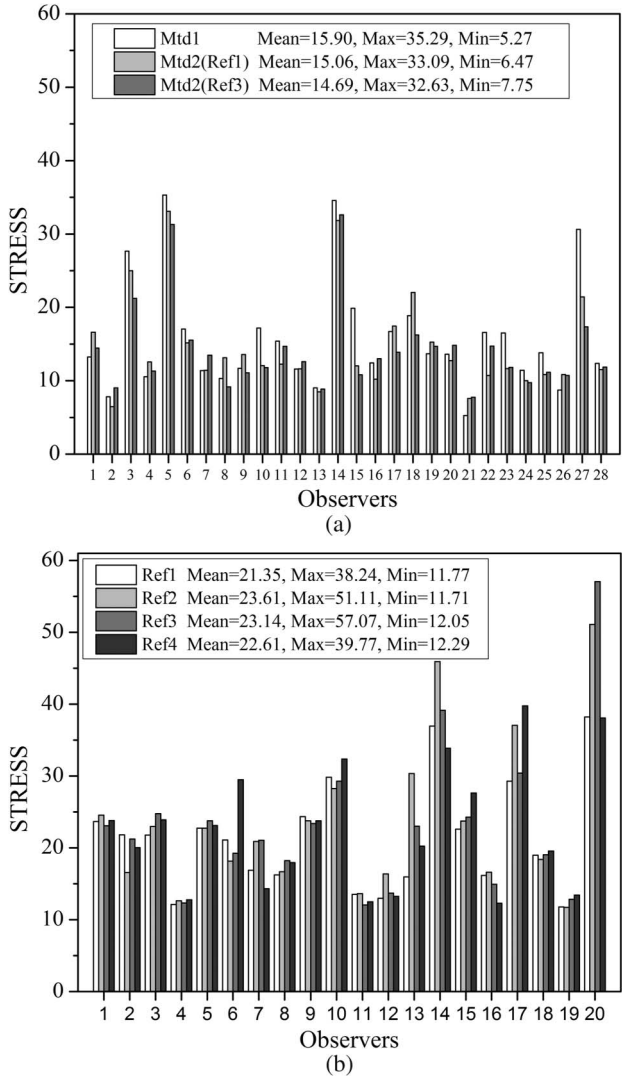


Fig. 6. (a) Inter-observer variability results at NCSU. (b) Inter-observer variability results at RIT.

samples are Ref1:10R4/8, Ref2:10YR7/10, Ref3:10Y7/10, and Ref4:10B3/6. The results shown in Table 4 show that no significant differences in perceived saturation based on methods were obtained.

The STRESS between visual results obtained from two methods was also calculated and was 4.16 for Mtd1 against Mtd2Ref1, 5.30 for Mtd1 against Mtd2Ref3, and 5.30 for Mtd2Ref1 against Mtd2Ref3, indicating again that no significant difference in responses based on methods is evident.

Comparable analyses for the results in the RIT experiment are shown in Table 5. In this case the average SD values for the four references used are: 0.71, 1.07, 1.07, and 0.78, and the average SE values are: 0.09, 0.14, 0.14, and 0.10, respectively. These values are almost twice those obtained in the NCSU experiment. This may be due to a slight difference in the experimental procedure used in two locations and/or related to the two different sets of observers or number of observers. The SD and SE of results obtained from Ref2 (10YR7/10) and Ref3 (10Y7/10) are found to be larger than those based on Ref1 (10R4/8) and Ref4 (10B3/6). Ideally, the values in the diagonal direction in Table 4 should be 1. Of the results shown, however, only sample 2 (10R4/8) has a value close to 1. This

**Table 2. Mean Intra-observer Variability Results of Three Trials at NCSU and RIT, Based on STRESS**

| | NCSU | | | RIT | | | |
|---|---|---|---|---|---|---|---|
| | Mtd1 | Mtd2(Ref1) | Mtd2(Ref3) | Ref1 | Ref2 | Ref3 | Ref4 |
| Mean | 12.05 | 10.53 | 10.47 | 17.93 | 18.73 | 18.63 | 18.60 |
| Max | 23.91 | 20.84 | 20.00 | 40.72 | 46.06 | 39.35 | 52.35 |
| Min | 3.26 | 2.94 | 2.74 | 6.89 | 8.15 | 6.84 | 7.36 |

**Table 3. Mean, SE, and SD of Saturation Determined Visually in Experiments Conducted at NCSU and RIT**

| | | | 10R | 10YR | 10Y | 10GY | 10G | 10BG | 10B | 10PB | 10P | 10RP | MEAN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NCSU | SD | Mean | 0.32 | 0.45 | 0.27 | 0.32 | 0.31 | 0.32 | 0.43 | 0.26 | 0.27 | 0.30 | 0.32 |
| | | Max | 0.63 | 0.74 | 0.48 | 0.50 | 0.47 | 0.41 | 0.77 | 0.31 | 0.41 | 0.59 | 0.53 |
| | | Min | 0.14 | 0.20 | 0.14 | 0.19 | 0.18 | 0.21 | 0.18 | 0.19 | 0.18 | 0.17 | 0.18 |
| | SE | Mean | 0.03 | 0.05 | 0.03 | 0.03 | 0.03 | 0.03 | 0.05 | 0.03 | 0.03 | 0.03 | 0.03 |
| | | Max | 0.07 | 0.08 | 0.05 | 0.05 | 0.05 | 0.04 | 0.08 | 0.03 | 0.04 | 0.06 | 0.06 |
| | | Min | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| RIT | SD | Mean | 0.75 | 0.86 | 0.70 | 0.83 | 1.01 | 0.99 | 0.94 | 1.09 | 1.04 | 0.87 | 0.91 |
| | | Max | 1.25 | 1.44 | 0.98 | 1.12 | 1.24 | 1.23 | 1.38 | 1.29 | 1.31 | 1.34 | 1.26 |
| | | Min | 0.34 | 0.36 | 0.40 | 0.60 | 0.70 | 0.77 | 0.47 | 0.86 | 0.80 | 0.57 | 0.58 |
| | SE | Mean | 0.10 | 0.11 | 0.09 | 0.11 | 0.13 | 0.13 | 0.12 | 0.14 | 0.13 | 0.11 | 0.12 |
| | | Max | 0.16 | 0.19 | 0.13 | 0.14 | 0.16 | 0.16 | 0.18 | 0.17 | 0.17 | 0.17 | 0.16 |
| | | Min | 0.04 | 0.05 | 0.05 | 0.08 | 0.09 | 0.10 | 0.06 | 0.11 | 0.10 | 0.07 | 0.08 |

indicates a possible visual interactive effect among the references employed in the perceptual assessment of saturation.

A similar trend to that seen in the NCSU experiment was also observed for the RIT experiment, i.e., the variability increases with an increase in saturation or chroma of samples with the same hue. However, the hue dependent variability is not as obvious as that noted in the NCSU experiment when using reference 1 or 4. Nonetheless, the variability for samples 10G, 10BG, 10B, 10PB, and 10P was found to be larger than that for other hues. The mean visual saturation values from RIT are shown in Table 5.

A comparison of the calculated STRESS based on different methods indicates variability in the mean visual saturation based on different references. The results based on reference 4 are largely different from those based on other references. The STRESS for reference 1 against 4 is 6.55, that between references 2 and 4 is 8.35 and that for references 3 and 4 is 7.74. These results are an indicator of the complexity of assessing saturation visually.

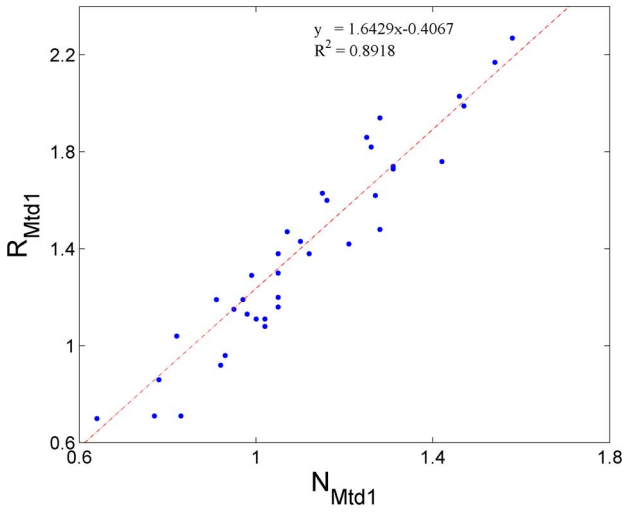


Fig. 7. Agreement between NMtd1 and RMtd1 results.

### C. Comparison of Results from Two Experiments

The NCSU experimental results using method 1 and method 2 are denoted as NMtd1 and NMtd2. The grand mean visual saturation data for four references from the RIT experiment are denoted RMtd1 and those for references 1 and 3, which is the same as that used in the NCSU experiment, are denoted RMtd2. Scatter plots of NMtd1 against RMtd1, and NMtd2 against RMtd2 are shown in Figs. 7 and 8.

**Table 4. Perceived Saturation Results Based on Various Methods**

| | 10R4/8 | 10YR7/10 | 10Y7/10 | 10B3/6 |
|---|---|---|---|---|
| Mtd1 | 0.96 | 1.03 | 1.01 | 1.00 |
| Mtd2Ref1 | 1.04 | 1.04 | 1.09 | 1.06 |
| Mtd2Ref3 | 1.02 | 1.08 | 1.04 | 1.09 |

**Table 5. Mean Visual Saturation of Reference Samples Based on Different References Determined at RIT**

| | Visual Saturation Based on Each Reference | | | |
|---|---|---|---|---|
| Sample ID | 10R4/8 | 10YR7/10 | 10Y7/10 | 10B3/6 |
| 10R4/8 | **1.08** | 1.15 | 1.09 | 1.11 |
| 10YR7/10 | 1.15 | **1.14** | 1.16 | 1.19 |
| 10Y7/10 | 1.35 | 1.42 | **1.28** | 1.47 |
| 10B3/6 | 1.29 | 1.46 | 1.26 | **1.21** |

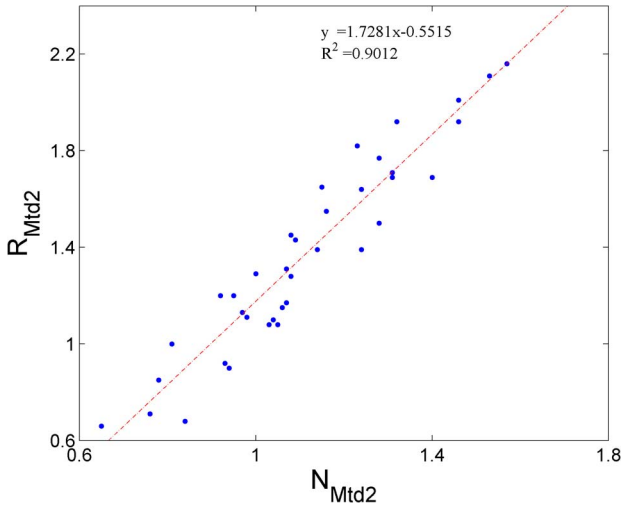


Fig. 8. Agreement between NMtd2 and RMtd2 results.

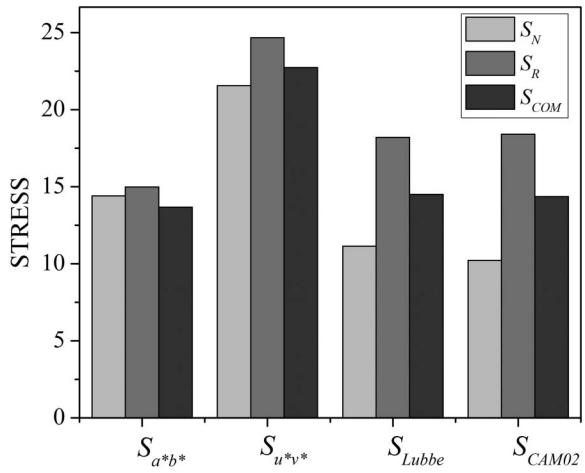


Fig. 9. Comparison of STRESS for different saturation models against NCSU, RIT, and combined experimental results.

The two sets of results show general agreement with a STRESS of 10.92 for Mtd1 in two locations and 11.79 for NMtd2 against RMtd2. The results are strongly related with a correlation coefficient of 0.95. However, there are also certain differences between results:

(1) The visual saturation range of the two datasets is different. The ranges for NMtd1 and NMtd2 are 0.61 to 1.59 and 0.65 to 1.57, respectively, and those for RMtd1 and RMtd2 are 0.70 to 2.27, and 0.66 to 2.16, respectively.

(2) Visual steps between two successive chroma values, especially for samples with high chroma, are larger in the RIT experiment. This is in part due to responses from an observer with large visual ratings.

### D. Performance of Various Saturation Models

The performance of four saturation models, i.e., $S^*_{ab}$, $S^*_{uv}$, $S_{\text{Lübbe}}$, and $S_{\text{CAM02}}$, against visual data were tested using the STRESS index. Models' performances against visual saturation were compared for Mtd1, Mtd2Ref1, Mtd2Ref3, as well as the NCSU experimental grand mean (denoted $S_N$), as shown in Table 6 and Fig. 9. Due to differences in the scales, results were normalized to a range of 0–10 for comparison. The normalization does not affect the STRESS values.

Results in Table 6 indicate that $S_{\text{CAM02}}$ outperformed all other models for the NCSU data with a mean STRESS value of 10.22. The Lübbe formula resulted in a slightly higher STRESS value than $S_{\text{CAM02}}$. Both significantly outperformed the $S^*_{ab}$, and $S^*_{uv}$ models.

For the RIT experiment, $S^*_{ab}$ gives the best performance, with a STRESS of 15.0. $S_{\text{Lübbe}}$ and $S_{\text{CAM02}}$ models gave comparable performances, with STRESS values of approximately 18.3. The worst results for both mean sets of experimental data were obtained from $S^*_{uv}$.

For the combined NCSU and RIT data the mathematical mean was calculated. STRESS between combined data, denoted $S_{\text{COM}}$, and that computed by four models are also shown in Table 6. For the NCSU data $S_{\text{CAM02}}$ statistically outperformed the other three models, and $S_{\text{Lübbe}}$ was found to be significantly better than $S^*_{ab}$ and $S^*_{uv}$. For the RIT data, $S^*_{ab}$ performed statistically better than $S^*_{uv}$. For the combined dataset, modeling by $S^*_{ab}$, $S_{\text{Lübbe}}$, and $S_{\text{CAM02}}$ gave comparable results, with no significant difference between models; however, $S^*_{uv}$ performed significantly worse than all other models.

**Table 6. STRESS between Perceived Saturation Against Saturation Based on Various Models (Bold Letter Indicates Models with the Best Agreement)**

| | | STRESS | | | |
|---|---|---|---|---|---|
| Location/Method | | $S^*_{ab}$ | $S^*_{uv}$ | $S_{\text{Lübbe}}$ | $S_{\text{CAM02}}$ |
| NCSU | Mtd1 | 14.38 | 21.60 | 11.87 | **11.19** |
| | Mtd2Ref1 | 13.84 | 21.15 | 10.67 | **9.65** |
| | Mtd2Ref3 | 15.61 | 22.22 | 11.14 | **10.00** |
| | **$S_N$** | 14.41 | 21.56 | 11.15 | **10.22** |
| RIT | Ref1 | **14.31** | 23.93 | 18.10 | 18.39 |
| | Ref2 | **16.46** | 25.44 | 19.63 | 19.87 |
| | Ref3 | **15.78** | 25.58 | 18.10 | 18.43 |
| | Ref4 | **15.43** | 24.97 | 18.45 | 18.33 |
| | **$S_R$** | **14.99** | 24.68 | 18.20 | 18.41 |
| COMBINED DATA | | **13.67** | 22.73 | 14.51 | 14.36 |

## 4. CONCLUSIONS

Two psychophysical experiments were conducted at two laboratories to collect new saturation judgments using similar methodology and the same Munsell atlas samples. STRESS values indicate that the average inter- and intra-observer variability of the experimental results at NCSU is smaller than or comparable to that of typical color difference evaluation experiments. Results for the RIT experiment are slightly larger. For samples with higher chroma values, responses were found to be less consistent, reflected by higher associated SE. The variability was also found to be hue dependent.

Visual results within each experiment for different methods or different reference samples show good agreement, with an average STRESS of 4.92 for the NCSU experiment and 5.52 for the RIT experiment. The STRESS when comparing results from two locations is 11.36 indicating reasonable agreement between the two sets of data.

Among the four saturation models, i.e., $S^*_{ab}$, $S^*_{uv}$, $S_{\text{Lübbe}}$, and $S_{\text{CAM02}}$ tested against the visual responses obtained. $S_{\text{CAM02}}$ performed best for the NCSU results, slightly better than the Lübbe model. Both significantly outperformed $S^*_{ab}$ and $S^*_{uv}$. For the RIT data, $S^*_{ab}$ was found to be the best, followed by the $S_{\text{Lübbe}}$ and $S_{\text{CAM02}}$ models, with $S^*_{uv}$ being the worst.

According to STRESS, for the combined dataset, $S^*_{ab}$, $S_{\text{Lübbe}}$, and $S_{\text{CAM02}}$ resulted in similar performance, better than $S^*_{uv}$.

Given the differences in means and ranges of STRESS values of essentially identical experimental conditions (Ref1 and Ref3) it appears that a large number aof observers (about 50) is required to establish statistically solid mean perceptual saturation data.

## APPENDIX A

**Table 7. Colorimetric Values and the Visual Saturation of Samples and References (bold italic)[a]**

| Sample ID | Munsell | $X_{10}$ | $Y_{10}$ | $Z_{10}$ | $S_N$ | SD | SE | $S_R$ | SD | SE |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10R 4/6 | 14.45 | 11.31 | 6.24 | 0.82 | 0.15 | 0.02 | 0.71 | 0.34 | 0.044 |
| 2 | ***10R 4/8*** | 15.85 | 11.28 | 4.12 | 0.99 | 0.14 | 0.02 | 1.11 | 0.56 | 0.072 |
| 3 | 10R 4/10 | 17.68 | 11.78 | 3.15 | 1.28 | 0.36 | 0.04 | 1.48 | 0.87 | 0.113 |
| 4 | 10R 4/12 | 18.11 | 11.32 | 1.85 | 1.48 | 0.63 | 0.07 | 1.99 | 1.27 | 0.164 |
| 5 | 10YR 7/8 | 43.39 | 40.15 | 13.38 | 0.78 | 0.21 | 0.02 | 0.71 | 0.37 | 0.047 |
| 6 | ***10YR 7/10*** | 45.99 | 41.75 | 9.27 | 1.04 | 0.23 | 0.03 | 1.16 | 0.55 | 0.071 |
| 7 | 10YR 7/12 | 45.63 | 40.43 | 5.35 | 1.44 | 0.61 | 0.06 | 1.76 | 1.11 | 0.143 |
| 8 | 10YR 7/14 | 46.29 | 40.18 | 2.51 | 1.58 | 0.74 | 0.08 | 2.27 | 1.45 | 0.188 |
| 9 | 10Y 7/6 | 37.71 | 40.95 | 16.31 | 0.63 | 0.22 | 0.02 | 0.70 | 0.84 | 0.108 |
| 10 | 10Y 7/8 | 38.02 | 41.16 | 10.97 | 0.78 | 0.20 | 0.02 | 0.86 | 0.41 | 0.052 |
| 11 | ***10Y 7/10*** | 37.48 | 40.55 | 6.17 | 1.04 | 0.16 | 0.02 | 1.38 | 0.75 | 0.096 |
| 12 | 10Y 7/12 | 36.92 | 39.99 | 3.07 | 1.32 | 0.48 | 0.05 | 1.74 | 0.92 | 0.119 |
| 13 | 10GY 5/6 | 14.75 | 20.06 | 10.94 | 0.82 | 0.19 | 0.02 | 1.04 | 0.67 | 0.086 |
| 14 | 10GY 5/8 | 12.86 | 19.45 | 7.48 | 0.98 | 0.26 | 0.03 | 1.19 | 0.72 | 0.093 |
| 15 | 10GY 5/10 | 11.85 | 19.63 | 5.39 | 1.29 | 0.50 | 0.05 | 1.62 | 1.15 | 0.149 |
| 16 | 10G 4/6 | 7.83 | 12.36 | 11.63 | 0.98 | 0.18 | 0.02 | 1.13 | 0.75 | 0.097 |
| 17 | 10G 4/8 | 7.09 | 12.88 | 11.45 | 1.1 | 0.28 | 0.03 | 1.43 | 1.15 | 0.149 |
| 18 | 10G 4/10 | 6.09 | 13.12 | 11.25 | 1.26 | 0.47 | 0.05 | 1.86 | 1.24 | 0.160 |
| 19 | 10BG 3/4 | 4.91 | 6.88 | 9.99 | 1.01 | 0.21 | 0.02 | 1.08 | 0.84 | 0.108 |
| 20 | 10BG 3/6 | 4.32 | 7.02 | 11.26 | 1.2 | 0.34 | 0.04 | 1.42 | 1.03 | 0.133 |
| 21 | 10BG 3/8 | 4.00 | 7.71 | 13.40 | 1.25 | 0.41 | 0.04 | 1.82 | 1.26 | 0.162 |
| 22 | 10B 3/4 | 5.94 | 7.17 | 13.28 | 0.91 | 0.21 | 0.02 | 0.92 | 0.49 | 0.063 |
| 23 | ***10B 3/6*** | 5.83 | 7.43 | 16.87 | 1.04 | 0.20 | 0.02 | 1.3 | 0.93 | 0.121 |
| 24 | 10B 3/8 | 5.47 | 7.68 | 19.89 | 1.31 | 0.55 | 0.06 | 1.73 | 1.06 | 0.137 |
| 25 | 10B 3/10 | 5.53 | 8.40 | 24.65 | 1.55 | 0.77 | 0.08 | 2.17 | 1.39 | 0.180 |
| 26 | 10PB 3/4 | 7.06 | 6.77 | 12.97 | 0.9 | 0.23 | 0.02 | 1.19 | 0.92 | 0.119 |
| 27 | 10PB 3/6 | 7.63 | 6.94 | 16.07 | 0.99 | 0.20 | 0.02 | 1.29 | 1.22 | 0.158 |
| 28 | 10PB 3/8 | 8.68 | 7.49 | 20.49 | 1.12 | 0.29 | 0.03 | 1.38 | 1.08 | 0.139 |
| 29 | 10PB 3/10 | 9.00 | 7.39 | 23.92 | 1.16 | 0.31 | 0.03 | 1.6 | 1.30 | 0.167 |
| 30 | 10P 3/4 | 7.66 | 6.52 | 10.00 | 0.94 | 0.22 | 0.02 | 1.15 | 0.89 | 0.115 |
| 31 | 10P 3/6 | 8.68 | 6.74 | 12.13 | 1.06 | 0.18 | 0.02 | 1.47 | 1.09 | 0.140 |
| 32 | 10P 3/8 | 9.38 | 6.67 | 13.62 | 1.15 | 0.29 | 0.03 | 1.63 | 1.09 | 0.141 |
| 33 | 10P 3/10 | 10.16 | 6.56 | 15.01 | 1.26 | 0.41 | 0.04 | 1.94 | 1.35 | 0.174 |
| 34 | 10RP 3/4 | 8.40 | 6.73 | 6.78 | 0.92 | 0.21 | 0.02 | 0.96 | 0.63 | 0.081 |
| 35 | 10RP 3/6 | 8.78 | 6.35 | 6.16 | 1.05 | 0.20 | 0.02 | 1.20 | 0.85 | 0.110 |
| 36 | 10RP 3/8 | 9.49 | 6.75 | 6.44 | 1.02 | 0.20 | 0.02 | 1.11 | 0.73 | 0.094 |
| 37 | 10RP 3/10 | 11.03 | 6.29 | 6.00 | 1.45 | 0.59 | 0.06 | 2.03 | 1.39 | 0.179 |

[a]Illuminant D65, CIE 1964 standard observer, $S_N$ and $S_R$ are the grand mean visual saturation from the NCSU and RIT experiments, SD, and SE represent the mean SD and SE.

## APPENDIX B

### 1. Procedure for Visual Assessment of Saturation

This experiment aims to elucidate our understanding for the perception of the term "saturation." There are three sections in this experiment.

### A. Section I

In this section the aim is to provide an understanding of the meaning of the term saturation by showing a set of samples on a calibrated monitor. First, a set of samples with given saturation values (0, 1, 3, 5, 7), but different hue and lightness, are displayed. Second, samples with the same hue (Red, Yellow, Green, and Blue), but different saturation and lightness values are shown. Finally, samples with the same lightness (1, 3, 5), but different saturation and hue are shown. The arrangement of samples will also be explained. This process can be repeated during the experiment if the observer is not certain about their understanding of saturation.

### B. Section II

In this section, the observer is asked to determine a numerical value for the saturation of 37 samples. For each sample, four different assessments involving four different reference samples are conducted. Observers will wear a gray lab coat and gray gloves and sit in front of the empty viewing booth for at least two minutes to adapt to the source. During the experiment, a test sample and a reference sample will be placed on a custom stand at a 45° viewing angle, with a gap between them.

An arbitrary value of 1 is given to the reference sample on the left, and the observer gives a numerical rating of the saturation of the test sample based on the reference. The value can be multiples or fractions of 1, e.g., 0.5, 1.2 or 2 or more.

### C. Section III

In this section, the observer will assess the saturation of 37 samples in the presence of four reference samples shown simultaneously. The assigned numerical saturation value of all reference samples is 1, and the observer will give a rating for the test sample based on the reference samples. The rating can be multiples or fractions of 1, e.g., 0.5, or 2 or more.

**Notes:**

- Observers are notified that there are no right or wrong answers.
- If they find it difficult to provide a rating for the saturation of samples during the experiment, they may ask for additional training.
- Observers are asked to refrain from handling the samples and ask the experimenter if they would like to move them.

### 2. Saturation

Thank you for participating in our experiment to measure our perceptions of **saturation**.

**Saturation** is one attribute of our perception of color. Other attributes include lightness (black is of low lightness, white is of high lightness) and hue (often described by color names such as red, yellow, green, blue). For this experiment, we are interested in the perception of **saturation** independent of perceived lightness or hue.

The formal, technical, definition of **saturation** is:

*the colorfulness of an area judged in proportion to its brightness*,

where colorfulness is: *the attribute of a visual perception according to which the perceived color of an area appears to be more or less chromatic.*

And brightness is: *the attribute of a visual perception according to which an area appears to emit, or reflect, more or less light.*

More practically, saturation can be thought of as how much a color stimulus differs from a neutral (white, gray, or black) stimulus in terms of the intensity of perceived hue present. A neutral, or gray, color has no hue present and therefore a saturation value of zero. A vivid red color is clearly different from gray in that it has a hue with an intensity and therefore has a saturation significantly greater than zero [the exact amount will be defined by the reference color(s) in the experiment].

You are being shown examples of sets of colors of constant saturation at various lightness levels. The different sets are for various hues and saturation levels. Each set is of constant saturation and labeled for the saturation to provide an idea of what changes in saturation look like. It is most important to note that each set of color samples illustrates constant saturation across a range of lightness rather than any changes in saturation.

There are other ways to describe the intensity of hue in a color stimulus. These are known as colorfulness and chroma. In this experiment, we are not interested in those attributes. We are only interested in your perception of saturation.

If you need any clarification on the definition of saturation, please ask the experimenter to review the examples with you. Specific instructions for defining the reference color(s) and completing the experiment follow.

Thank you.

## ACKNOWLEDGMENTS

The authors thank Mr. Art Schmehling, Munsell Color Services Business Manager (X-Rite) for donation of Munsell sheets. Thanks are also due to all observers who took part in the study.

## REFERENCES

1. R. W. G. Hunt, "The specification of colour appearance. I. Concepts and terms," Color Res. Appl. **2**, 55–68 (1977).
2. I. Newton, *Opticks* (Smith and Walford, 1704), p. 117.
3. H. v. Helmholtz, *Handbuch der Physiologischen Optik* (Leopold Voss, 1867), p. 283.
4. M. Richter, I. Schmidt, and A. Dresler, *Grundriss der Farbenlehre der Gegenwart* (Steinkopff, 1940).
5. D. L. MacAdam, "Projective transformations of I. C. I. color specifications," J. Opt. Soc. Am. **27**, 294–297 (1937).
6. D. B. Judd, "A Maxwell triangle yielding uniform chromaticity scales," J. Opt. Soc. Am. **25**, 24–35 (1935).
7. CIE, "A colour appearance model for colour management systems: CIECAM02," CIE Publication 159 (CIE Central Bureau, 2004).
8. E. Lübbe, "Sättigung im CIELAB-Farbsystem und LSh-Farbsystem," Ph.D. dissertation (Technische Universität Ilmenau, 2011).
9. L. Y. G. Juan and M. R. Luo, "Magnitude estimation for scaling saturation," Proc. SPIE **4421**, 575–578 (2002).
10. http://www.vcsconsulting.co.uk/home.html, retrieved 4/8/2014.
11. P. A. Garcia, R. Huertas, M. Melgosa, and G. Cui, "Measurement of the relationship between perceived and computed color differences," J. Opt. Soc. Am. A **24**, 1823–1829 (2007).
12. M. Melgosa, P. A. García, L. Gómez-Robledo, R. Shamey, D. Hinks, G. Cui, and M. R. Luo, "Notes on the application of the standardized residual sum of squares index for the assessment of intra- and inter-observer variability in color-difference experiments," J. Opt. Soc. Am. A **28**, 949–953 (2011).
13. http://danielsoper.com/statcalc3/calc.aspx?id=, retrieved 4/8/2014.
14. E. Kirchner and N. Dekker, "Performance measures of color-difference equations: correlation coefficient versus standardized residual sum of squares," J. Opt. Soc. Am. A **28**, 1841–1848 (2011).
15. M. R. Luo and B. Rigg, "Chromaticity-discrimination ellipses for surface colours," Color Res. Appl. **11**, 25–42 (1986).
16. R. Berns, D. H. Alman, L. Reniff, G. D. Snyder, and M. R. Balonon-Rosen, "Visual determination of suprathreshold color-difference tolerances using probit analysis," Color Res. Appl. **16**, 297–316 (1991).