

Methodology for constructing a colour-difference acceptability scale

Baptiste Laborie^{1,2}, Françoise Viénot², Sabine Langlois¹

¹ Renault, Direction de la Recherche, 1 Avenue du Golf, 78288, Guyancourt Cedex, France

² Muséum national d'histoire naturelle, Centre de Recherche sur la Conservation des Collections, MNHN-CNRS-MCC, Paris, France

Abstract

Observers were invited to report their degree of satisfaction on a 6-point semantic scale with respect to the conformity of a test colour with a white reference colour, simultaneously presented on a PDP display. Eight test patches were chosen along each of $+a^*$, $-a^*$, $+b^*$, $-b^*$ axes of the CIELAB chromaticity plane, at $Y=80 \pm 2$ cd.m^2 . Experimental conditions reliably represented the automotive environment (patch size, angular distance between patches) and observers could move their head and eyes freely.

We have compared several methods of category scaling, the Torgerson-DMT (Torgerson, 1958) method, two versions of the regression method (Bonnet's (Bonnet, 1986) and logistic regression) and the medians method. We describe in detail a case where all methods yield substantial but slightly different results. The solution proposed by the regression method which works with incomplete matrix and yields results directly on a colorimetric scale is probably the most useful in this industrial context. Finally we summarize the implementation of the logistic regression method over four hues and for one experimental condition.

Keywords

Acceptability, colour-difference, category scaling, plasma display panel.

Contact details

Baptiste Laborie, laborie@mnhn.fr

Shortened title

Colour-difference acceptability scale

Introduction

Purpose

Car manufacturers are increasingly incorporating displays inside their vehicles. Due to the increasing amount of displayed information and to the large variety of technologies, one of the main issues raised by these new interfaces is the conformity of the screens with a target and specifically in terms of displayed colours. Colorimetry proposes different solutions in terms of colour difference perceptibility and colour-difference acceptability (Witt, 2007). Colour-difference perceptibility deals with the ability of an observer to detect a difference between two colour samples. The smallest perceptible difference is known as a just-noticeable difference (JND). In a psychophysics experiment where a series of colour tests are to be compared with a colour reference, a colour-difference threshold can be derived from the statistical count of non-perceived and perceived colour differences. However, practically, in an industrial process, repeatability and reproducibility of a colour target may be difficult to achieve at a reasonable cost. Moreover, when asked to provide a visual judgment, an observer may interpret the acceptable colour-difference, depending upon the intended or anticipated end use of the product (Judd, 1975). Thus colour-difference acceptability results from a compromise between the process outcome and the customer expectations (Berns, 2000). An additional complexity is that the user environment may vary in terms of stimulus configuration, background and luminous adaptation. To characterize displays in the real automotive context, information is needed on colour comparison experiments when the background is not uniform, when the stimuli are positioned at a distance, when the observer can freely explore the visual field.

Background

When a producer must deliver a product of the desired colour with its variation controlled to an extent appropriate for the product's use and the customer's expectations, he may start with the colour-difference equations recommended by the CIE for specifying industrial colour-differences. The CIE ΔE_{94} and CIEDE2000 equations are based on CIELAB colour difference specifications and can be used to derive-colour difference tolerances by introducing factors to correct for the parametric

effects of various conditions of use (CIE, 1995; CIE, 2001). Meanwhile, rigorous reference conditions define material and viewing environment characteristics to which the colour-difference model applies. Unfortunately, several viewing conditions imposed by, such as uniform background, sample pairs with direct edge contact and foveal viewing, may considerably deviate from the reference viewing conditions in practical situations. Since the CIECAM02 colour appearance model has been developed to predict corresponding colours in asymmetric conditions, allowing for surround and background effects, a few attempts have been made to propose modified versions of CIECAM02 to fit small colour difference (SCD) and large colour difference (LCD) data sets. The LCD group includes six data sets having CIELAB colour difference values ranging from 9 to 14 with an average of 10. Thus a unique polar structure consisting of lightness, colourfulness and hue angle could be used to develop uniform colour spaces in which equal distances approximately represent equal colour differences (Luo, Cui, Li, 2006). Whereas a large colour-difference is easily perceived, it seems difficult for an observer to identify the colour attribute responsible for a colour difference. Only 72.4% successful identification was obtained for an average value of 15.8 CIELAB units (Melgosa et al., 2000).

Usually, every manufacturing company defines experimentally its own colour tolerance. Thus, several observers are invited to judge whether the colour difference between a sample and standard is acceptable for a given set of viewing conditions or with respect to the end-use of the product. Pairs of samples are sorted between “pass” and “fail”. Cumulative percentages for each class are plotted versus the instrumental colour difference calculated using a colour-difference equation. A decision is made that minimizes the number of instrumental wrong decisions (Berns, 2000). The procedure is very efficient economically. Nonetheless, it needs to be repeated every time the quality of the manufacturing process changes.

The hypothesis that colour discriminability and colour appearance are controlled by a common set of mechanism has been tested through experimental determination of JNDs and asymmetric matching. The proposed model presumes that the JND proportion has equal changes in the neural response along a single stimulus dimension

and therefore reflect the local slope of an appearance response function. Thus discrimination data can be used to infer the appearance response function and eventually using a parametric description of this function (Le Grand, 1972; Hillis, Brainard, 2005). Whichever generality the model has, it satisfies only the limited set of stimulus conditions chosen in the experiment.

An additional difficulty arises when two patches that are being compared are well apart, because peripheral colour vision is inevitably implicated. Colour discrimination in the periphery of the visual field has been examined through various psychophysical approaches. Asymmetric colour matches between a foveal three primary colour mixture and an extrafoveal monochromatic test showed a progressive reduction in size and shape of the chromaticity diagram with increasing distance from the fovea up to 50° (Moreland, 1972). Discrimination that depends only upon S-cones was improved by introducing a small gap between the two fields to be compared (Boynton, Hayhoe & Macleod, 1977). Evaluation of the effect of sample proximity upon threshold colour differences and upon sensitivity to small but clearly perceived colour differences indicated that field separation affects chromatic discrimination while no measurable loss of supra-threshold chromatic discrimination was recorded when the test field and the comparison field were separated by as much as 4.1° of visual angle. Nevertheless, large observer variability was encountered. (Sharpe and Wyszecki, 1976). Measured by a method of two-alternative spatial forced choice along either the $L/(L+M)$ or the $S/(L+M)$ axis of colour space, chromatic discrimination was found optimal when there was a small spatial interval between the boundaries of the stimuli; thereafter thresholds rose moderately with increasing angular separation, up to 10° . The two stimuli were presented shortly (100 ms) and at 5° eccentricity with a fixation point (Danilova and Mollon; 2006). Provided that the stimulus size is optimal (8°), colour stimuli along the $(S-(L+M))$ or the $(L-M)$ chromatic direction could be reliably detected, identified and discriminated at eccentricities up to 50° . Although, the decline in reddish-greenish $L-M$ sensitivity was greater than for bluish-yellowish $(S-(L+M))$ sensitivity, the decrease in sensitivity with peripheral presentation could be compensated by increasing the size of the stimulus (Hansen, Pracejus, & Gegenfurtner, 2009). Visual experiments were conducted to investigate parametric effects of sample separation and sample size in

assessing colour difference. The observer was asked to grade the colour difference between two samples with respect to the differences between a series of samples from a grey scale and a “standard” grey sample. It was found that if both grey scale pair and test pair have a 3-in gap (3-in is also the size of a sample), the colour difference perceived is slightly smaller (8%) than if both pairs are in hairline separation. It is clear that, in this condition, the effect of sample separation is reduced out by the same separation of the grey scale pair (Xin, Lam, Luo, 2001).

Proposal

Finally, the user satisfaction is the main concern for manufacturers. What is the acceptable colour-difference for most observers? We acknowledge that the measurement of colour difference threshold is not of any help when dealing with suprathreshold colour differences. We have to design an experiment to measure the acceptability of colour differences. Due to the specific use of colour in displays, the measurement should be made in similar conditions to the automotive context. Furthermore, because of the ongoing change in technology, a unique decision point might not be exploited in the future, so a category scale is preferred.

To answer the question “What is the acceptable colour-difference for most observers?”, we propose to ask a number of observers what would be their degree of satisfaction in terms of colour conformity if they were offered to compare the colour of two displays mounted on the dashboard of the vehicle. Colours should be presented to the observer as far as possible in real life conditions where the observer explores the stimulus at will. We propose to construct an acceptability scale by linking the CIELAB colour-difference specification and the degree of satisfaction of the observers in terms of conformity of two colours. Category scaling differs from threshold measurement in the sense that it deals with the subjective assessment of a perceived attribute which reflects the quality of a product rather than with the ability of the individual to discriminate between JND values of the attribute (Krantz, 1972; Engeldrum, 2000). In this study, we propose a category scale where category labels are adjectives easily understood by the observer. The underlying framework of a category scaling procedure is given by Torgerson as the Law of Categorical Judgment

(Torgerson, 1958, chapter 10). - “A psychological continuum of the attribute of interest is postulated”. Any given stimulus elicits a response in the psychological continuum of the subject. Nevertheless, the value of the response is not always the same and all values associated with this stimulus are normally distributed in the psychological continuum. - Another assumption is that the psychological continuum of the subject can be divided into a number of ordered categories. Additionally a given category boundary is not always located at a particular point on the continuum. It is defined by a mean location and dispersion. A complete form of the law of categorical judgment is usually too complex to bring a solution. For this reason, Torgerson has proposed a classification and simplifications of the problem. In this study, we have investigated the degree of satisfaction of observers in relation with the difference between two white patches using category scaling. The range of colour differences exceeds the traditional limits of colorimetric differences. We have chosen experimental conditions that reliably represent the automotive environment (patch size, angular distance between patches). Finally, we could compare the results from several scaling methods.

Methods

Display calibration

Colour patches were presented on a large Plasma Display Panel (PDP) (Pioneer KRP-500M, 50”, 1920 x 1080 pixels).

- Gamma setting.

Automatic controls were disconnected. “Brightness” and “Contrast” were fixed to avoid saturation. The final gamma equalled 2.2.

- RGB setting:

The maximum luminance of each primary was adjusted to obtain sRGB white ($x=0.3127$, $y=0.3290$). Contrary to what could be achieved in computer-controlled CRT displays (CIE, 1996, IEC, 1999), neither the standard matrix, nor the inter-channel matrix could provide a workable display calibration. The strategy adopted to circumvent this problem has been to build look-up-tables (LUT) around every target colour. Moreover, the displayed image is continuously changed by the built-in energy

saving mode to reduce its energy consumption. In particular, the light smoothly shuts down when no event occurs in the image. To avoid any instability of the luminance, we have refreshed at regular time intervals some part of the image that lied out of a region of interest.

Psychophysical experiment

1. Experimental conditions

This study was part of a research program on colour-difference acceptability in an automotive context. The experiment took place in the laboratory where the surrounding conditions were created to simulate the automotive interior. The observer was seated at a distance of 1 meter from a 50" PDP display. This geometrical configuration reproduces at best the geometry of a seated driver facing the steering wheel of the vehicle and seeing in his viewing field the dashboard with all displays. As a whole, it was possible to display colour patches at eccentricity as far as 40 degrees. The advantage of the PDP technology is its lambertian emissivity which ensures the validity of display calibration at any viewing angle. The observer could move his head and eyes freely. The background could be grey, black or the photograph of an automotive interior. Additionally, a video projector illuminated a part of the wall surface at the top of the display, at the same luminance as the background, in order to extend the field of view as through the windscreen. The photography (Figure 1) illustrates these experimental conditions.

A pair of square patches was displayed on the PDP display at a photopic luminance level, one being the reference colour, the other being the test colour.

- In this experiment, the reference colour was white (D65) at luminance $Y=80$ cd.m^2 .
- In this experiment, the background and the illuminated part of the wall were grey ($Y=73$ cd.m^2).
- The size of the two patches was 1 degree for simulating telltales or 5 degrees for simulating typical dashboard displays.

- An angular distance between the two patches was chosen so as to simulate the real automotive configuration. The angular distance between the two patches was either 10 arc min. between borders as a margin can never be avoided between two real displays, or one stimulus size between borders for simulating two non-contiguous real displays, or 30 degrees between centers that is approximately the distance between the straight ahead direction and a control display such as the global positioning system (GPS) panel, or 40 degrees between centers that is approximately the distance between the straight ahead direction and the most eccentric real advanced driving assistance system (ADAS) display.
- The time of presentation was controlled at 500 ms, 600 ms, 1,3 s, or 1.8 s, i.e. 500 ms plus at least two saccade durations between a pair of remembered positions (Hallett, 1986), in order to allow the observer to look to and from between the two patches according to the angular distance between them.

-2. Choice of the test colours

Eight test colours were chosen along the four hue axes $+a^*$, $-a^*$, $+b^*$, $-b^*$ of the CIELAB chromaticity plane, in the interval $[0, \Delta C^*_{\max}]$, at $Y=80 \pm 2 \text{ cd.m}^{-2}$. A colour pair is made of a white patch and a test colour patch. This makes 32 colour pairs for every size and angular distance experimental condition. Thus the null colour difference was included four times. Prior to the main experiment, three observers having experience with psychophysical experiments participated in the selection along each axis of the colour patch with maximum ΔC^* . They were given the same instructions as in the main experiment (see next paragraph) but an adaptive procedure allowed us to bracket the range of the stimulus around the boundary between “Very Unsatisfied” and “Unsatisfied” responses. It resulted that, in the main experiment, each experimental condition (size, angular distance, hue axis) was associated with a specific eight stimulus range that covers at best the psychometric scale of all observers (Table 1 gives an example).

3. Procedure

Each participant received instructions prior to the experiment. He was explained that he would have to rate his degree of satisfaction with respect to a difference of colour between two square patches. Instructions indicated that the square patches were

representative of two displays on two telltales. They required the participant to evaluate the difference of colour between the two colour patches as he would deliver a judgment about the conformity of the colour of two displays in an automotive context. Instructions (in French) were read to the observer prior the experiment: "Imagine that you have acquired a vehicle having two screens. The manufacturer wanted these screens to both the same white background in order to satisfy you. ... After having observed an image, we ask you to note your degree of satisfaction as for the respect of the intentions of the manufacturer."

The observer was invited to report his (her) degree of satisfaction on a 6-point semantic scale with respect to the similarity of the test colour with the reference colour. Category labels were "Very Unsatisfied", "Unsatisfied", "Lightly Unsatisfied", "Lightly Satisfied", "Satisfied", "Very Satisfied". The reason for choosing an even number psychometric scale instead of the common odd number practice arose from the car manufacturer requirement of a final decision on dissatisfaction / satisfaction. At the beginning of the experiment a training session was implemented for the observer to get used to the stimuli and the task. Then, three experimental cycles were conducted in a session, with a choice of three backgrounds. Only the results obtained on a grey uniform background have been used to construct the colour-difference acceptability scale presented here. Within each cycle, several runs were prepared, each corresponding to a [size, angular distance] pattern. The sequence of eight runs formed a digram-balanced design (Keppel and Wickens, 2004) to counterbalance both immediate sequential effects and the pairing of experimental conditions (size and angular distance) between subjects experiments. Within each run, the 32 (4 hues * 8 test colours) pairs of colour patches were randomized, among which four repetitions of the identity pair. We wrote our experiments in Matlab®, using the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997)."

4. Participants

Thirty-two subjects (eight women aged 20-29, eight women aged 30-40, eight men aged 20-29, eight men aged 30-40) participated in the experiment. All participants had

normal colour vision, tested with the Ishihara's and D15 tests for colour deficiency. All observers participated in all experiments.

Scaling methods

Scaling methods require every observer to rate each sample on a category scale, that is to place each sample in a category. Data are entered in a frequency matrix giving the number of individuals who placed a stimulus into a category. The frequency matrix has as many rows as the number of stimuli and as many columns as the number of categories (Table 1). The scaling methods that are described in the literature can be divided in two classes, depending whether one considers that the stimuli responses or the categories have fixed position on the psychological continuum of the observers.

Regression methods (1986)

Bonnet (1986) has introduced a method to establish a category scale under the assumption of fixed positions of the stimuli responses on the subjective continuum of the observer. The cumulative proportion matrix (Table 2) is calculated from the frequency responses matrix and transformed into a matrix of z -values from the [standard normal deviate](#) distribution. Then for each category in turn, we calculate the psychometric function that relates the z -values to the stimulus values (in physical units). The transformation in z -value linearizes the psychometric function. Using a linear regression model, we determine the equation of the function which has the form: $z_j = aS_j + b$, where S_j is the value of the stimulus i (in physical units) and z_j is the z -value associated with the stimulus for the relevant category. We solve the equation for $z_j = 0$ and obtain the category boundary, directly expressed in physical units (Table 3).

An akin method is to fit a sigmoid to all values of cumulative proportions (Figure 2), for each category. Then, both parameters of the sigmoid are determined using a logistic regression. The equation of the sigmoid allows us to calculate the limit of each category for 0.5 cumulative frequency, expressed in physical units (Table 3). The fitting quality can be evaluated by calculating the p -value of each parameter. The fitting is satisfactory if the p -value is smaller than 0.05.

Medians method

The medians method (Bonnet; 1986, page 148; Torgerson, 1958, page 72) is based on the assumption of fixed position of the categories on the subjective continuum of the observers. The category boundaries in physical units are given by the median value between categories. Practically, for each stimulus we calculate the cumulative frequency from the frequency responses matrix and allocate to each stimulus the scalar median value of the cumulative frequency. Then, plotting the scalar median value versus the stimulus physical value, we can determine by linear interpolation the physical value of the category boundaries. Note that taking the median value rather than the mean value reduces the impact of outliers.

Category scaling after Torgerson and DMT

Within Torgerson's framework (Torgerson, 1958), our experimental situation complies with the model of "Class II" which involves replications over individuals, each stimulus being judged once by each subject. The simplification corresponds to "Condition B" where it is assumed that the category boundaries are fixed throughout the experiment. The full category scaling consists of determining the scale values, their dispersion and the category boundaries. From the frequency matrix, a cumulative frequency matrix is obtained with a number of columns equal to the number of category boundaries, which yields a cumulative proportion matrix and is transformed in z -units. At the end, the procedure yields the sample scale values, their dispersion and the category boundaries on the same scale. Due to the matter we were dealing with, the frequency data matrix is likely to be incomplete. A solution has been proposed by Dieterich, Messick and Tucker that Engeldrum (2000) has named the DMT method and has implemented in MathCad®. The method gives the sample scale values and category boundaries on a z -scale. It also uses a weighting function that equals zero when the z -value is unity or zero values, and which is related to the variance of the proportion otherwise. As a minimum, there should be non-zero frequencies in at least three categories to yield a workable solution. We wrote the program in MatLab® to implement the solution. We added a step to transform the z -values into colorimetric specifications.

Results

Whereas results depend upon the experimental condition, we focus on methodology.

To facilitate the comparison of results obtained with different category scaling methods, we first propose an example related to colour differences along one hue axis, for one given experimental condition. Secondly, we summarize the results provided by the most direct scaling method over four hue axes and for the same experimental condition.

Example for the comparison of scaling methods

The example deals with the judgements of 32 observers. Indeed, not all methods would yield a complete solution when a small number of observers' results are entered. The case presented here refers to 1 degree patches, 30 degrees apart, on grey background and according to a variation along the $+a^*$ axis.

Original data are reported in Table 1. The first step consisting in the production of cumulative response functions is common to all methods (Table 2).

Table 3 presents the results obtained using the Bonnet's, logistic regression, medians and Torgerson-DMT methods and shows the estimated parameters provided by each method as well as the CIELAB values of the category boundaries.

- Bonnet's method

A linear regression is performed on z -values obtained from the cumulative proportion matrix. For every boundary, the slope a and intercept b stemming from the linear regression and the coefficient of determination of each of them is given. The values of all category boundaries are obtained except the "Very Satisfied / Satisfied" category boundary. Although the value is provided, it is negative (equal to -1.53) and has no physical reality.

- Logistic regression

A sigmoid is fitted on each column of the cumulative proportion matrix. The coefficients b_1 and b_2 are presented with the p - values (for each sigmoid). We notice that p -values always exceed 0.05. The category boundary “Very Satisfied / Satisfied” is negative (equal to -1.18) and has no physical reality. For this boundary, the p -values are larger than 0.51.

- Medians

The category boundaries are obtained except for the “Very Satisfied / Satisfied” boundary which cannot be calculated. Indeed we note that the rating of the identical colour pair is lower than the median rating that would have been attributed to this boundary.

- Torgerson-DMT

Torgerson-DMT method provides the values of all the boundaries as well as the position of the samples and their dispersion. Results are expressed in z -values (Table 4 and Figure 3). However, it is not possible to connect directly those values with the physical continuum. To establish the connection, an additional step (not included in the original Torgerson’s method) is necessary. Here, we have fitted a line to the distribution of z -values vs. CIELAB specification of samples and we have used this linear model to allocate CIELAB specification to the category boundaries.

- Comparison of the scaling methods

The comparison of the scaling methods shows that:

- No method has allowed us to obtain the category boundary “Satisfied / Very Satisfied”.

- The results returned by the four methods are only slightly different (Figure 4). Results from Bonnet's and logistic regression methods superimpose as expected since they only differ with respect to the fitting method.

- In the following section, as long as the difference of results for a given category boundary is small enough with respect to the car manufacturer expectation, we will present results from the logistic regression since the method directly exploits the observers' judgments.

Complete study of one experimental case over four hue axes

The results presented here (Table 5) refer to 1 degree patches, 30 degrees apart, on grey background. We report on the category boundaries obtained along the $-a^*$, $+a^*$, $-b^*$, $+b^*$ axes. For every category boundary, the colour tolerance is higher on the b^* axis than on the a^* axis, about twice as much higher. Such information is valuable for the car manufacturer. We observe that the category boundary "Satisfied / Very Satisfied" is never achieved along any axis. Computation has also shown that the p -values of the coefficients b_1 and b_2 specifying the "Satisfied / Very Satisfied" sigmoid are very high (0.40 to 0.83) and much higher than for the other category boundaries.

Discussion

Colour tolerance, Categorical scales and ecological validity

As stressed in the introduction, the measurement on colour difference threshold may not be of any help when dealing with suprathreshold colour differences, as it is often the case in industry. In the experimental case presented in the results section, which correspond to 1° patches 30° apart, the limit of satisfaction / dissatisfaction is a few CIELAB units, from 2.06 to 6.08, what is higher than the colour difference threshold but would not be considered as a "large colour difference" after Luo et al. (Luo, Cui, Li, 2006). Remarkably, the extent of the full category scales is large. This might be in

relation with the fact that it seems difficult for an observer to identify the colour attribute responsible for a colour difference, as reported by Melgosa et al. (2000). Above all, the full category scale is about twice as large for the $-b^* + b^*$ axis (4.92 and 6.08) than for the $-a^* + a^*$ axis (2.06 and 2.85). Thus neither the CIELAB metrics nor a metrics allowing for weighting independently the chroma and hue differences would render this asymmetry. Let us propose an explanation. It is possible that the subject's judgement is based on appearance. It is also possible that he is influenced by his everyday experience of the white colour mainly dominated by daylight. In such case, the larger tolerance judgement along the $-b^* + b^*$ would simply reflect the anisotropic distribution of natural light. The missing value for the lower category deserves attention. The difficulty in determining the limit "Very satisfied" / "Satisfied" comes from the low number of responses "Very satisfied", even for the identity patch. It should be emphasized that the experiment that we conducted is not a threshold experiment but an acceptability experiment. It is possible that the judgement of the observer rely on some interpretation.

We insist on the necessity to recruit a sufficient number of observers. When a large number of subjects is examined, and provided that the problem reduces to finding the boundary between "satisfied" and "unsatisfied" judgements, it is expected that all methods yield practically congruent information as all of them rest on the midpoint of the cumulative probability function where the data are robust and the function is usually close to linear. Precisely, the medians method merely retains data bracketing the midpoint, Bonnet's method excludes extreme data, and the logistic regression and DMT method reduce the weights of the extreme data of the full probability distribution. Although the DMT method is recognized as being the most elegant, it possesses drawbacks in practical situations such as the one that we have studied. In our example, all configurations have fortunately fulfilled the necessary completeness condition. Nevertheless, over the wide range of stimuli that were of interest to us, i.e. covering the full judgement scale from "Very Satisfied" to "Very Unsatisfied", extreme stimuli of the scale are likely to incompletely fill the full categories. For instance, if among all stimuli representative of the variety of items issued from different manufacturing processes, some are under the quality requirement of the client, missing categories could frequently occur. Then the completeness condition

might not be fulfilled. In the industrial context, the regression solutions are probably the most useful. They deal with sub-matrices of the frequency responses where the problem reduces to a threshold problem. Thus, when the exact values of the stimulus are known, the methods provide a straightforward positioning of the boundary.

Scaling large colour differences

The maximum likelihood difference scaling (MLDS) method was introduced for estimating suprathreshold differences between stimuli and assigning them numbers that accurately predict an observer's judgments (Maloney and Yang, 2003; Charrier et al., 2007). The series of stimuli that elicit a perceptual judgement should be distributed across a range of one-dimensional perceptual variables. For instance, the method was successfully applied to colour patches that fall on a greenish-reddish line in the colour space that is along the a^* direction. The results tell us whether the observer sees marked or slight change between adjacent stimuli. Such a piece of information would be valuable to a car manufacturer. However in the MLDS model, the authors have set the outmost stimulus scale values to 0 and 1, which prevents the user to compare scales in various directions of the colour space and to interpret the scale values in terms of a semantic scale. Indeed, what guarantees the elegance and the robustness of the MLDS method is, no semantic interpretation is given to the scale built by the observer. The results of the difference scaling do not tell us anything about the feeling of the observers whereas the industrial needs to relate human perception to a quantitative scale. We may add that the double pair arrangement required by the MLDS design would not be workable when the colour pair is made of well apart colour patches. Nevertheless, a comparison between MLDS approach and Torgerson-DMT or Bonnet's or logistic regression categorical scaling approach could be of the highest interest.

Conclusion and perspective

Finally, the colour-difference acceptability scales that we have constructed using the logistic regression method directly reflect the perception of the participants. They could not have been derived from a simple colour-difference threshold experiment. Designing the experiment with no repetition has shortened the experimental session

so that several observers could participate which, is of primary importance in an industrial and commercial context. In the future, we could compare homogeneous groups of observers selected among the participants. We could also extend the experiment to other colours centres and other backgrounds which would offer the possibility to investigate appearance and the role of cognitive processes in colour context.

Acknowledgements

We thank Xavier Chalandon (Renault) for encouraging this work, Jean Le Rohellec (CRCC), David Blumenthal (Renault) and Antoine Saint-Marcoux (Renault) for help in statistics, Kristyn Falkenstern for suggestions and observers for their availability.

All experiments were performed with Psychtoolbox-3 (<http://psychtoolbox.org>).

References

- Berns, R. (2000). *Billmeyer and Saltzman's Principles of color technology*, 3rd edition, Wiley, New York, USA.
- Bonnet, C. (1986). *Manuel pratique de psychophysique*, Armand Colin, Paris, France.
- Boynton, R. M., Hayhoe, M. M. and MacLeod, D. I. A. (1977). The gap effect: chromatic and achromatic visual discrimination as affected by field separation. *Optica Acta* 24, 159-177.
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision* 10, 433-436.
- Charrier, C., Maloney, L. T., Cherifi, H. and Knoblauch, K. (2007). Maximum likelihood difference scaling of image quality in compression-degraded images. *J. Opt. Soc. Am. A* 24, 3418-3426.
- CIE (1995). *Industrial Colour-Difference Evaluation*. CIE Publ. 116-1995.

CIE (1996). *The Relationship between Digital and Colorimetric Data for Computer-Controlled CRT Displays*. CIE Publ. 122-1996.

CIE (2001). *Improvement to Industrial Colour-Difference Evaluation*. CIE Publ. 142-2001.

Danilova, M. V. and Mollon, J. D. (2006). The comparison of spatially separated colours. *Vision Res.* 46, 823-36.

Engeldrum, P. (2000). *Psychometric Scaling: A Toolkit for Imaging Systems Development*. Imcotek Press, Winchester, USA.

Green, P, Colour Engineering Toolbox, (<http://www.digitalcolour.org/>).

Hallett, P. (1986). Eye Movements. In: *Handbook of Perception and Human Performance, Volume 1, Sensory Processes and Perception*. Wiley-Interscience, Toronto, Chapter 10, pp.1-112.

Hansen, T., Pracejus, L. and Gegenfurtner, K. R. (2009). Color perception in the intermediate periphery of the visual field. *Journal of Vision* 9, 1-12.

Hillis, J. M. and Brainard, D. H. (2005). Do common mechanisms of adaptation mediate color discrimination and appearance? Uniform backgrounds. *J. Opt. Soc. Am. A* 22, 2090-2106.

IEC (1999), *sRGB default RGB colour space*. International Electrotechnical Commission standard IEC 61966-2-1.

Judd, D. B. and Wyszecki, G. (1975). *Color in business, science, and industry*. Wiley, New York, p.40.

Keppel, G. and Wickens, T. (2004). *Design and analysis. A researcher's handbook* (4th Ed.), Pearson education, Upper Saddle River, New Jersey, USA.

Krantz, D. H. (1972). Visual scaling. In: *Visual Psychophysics, Vol. VII/4 of Handbook of Sensory Physiology* (eds. D. Jameson and L. M. Hurvich), Springer; New York, USA, pp. 660-689.

Le Grand, Y. (1972). *Optique physiologique, tome II: lumière et couleurs*. Masson, Paris, France. *Light, Colour and Vision*. English translation approved by Hunt, Walsch & Hunt, Chapman & Hall, London, 1968.

Luo M. R., Cui G. and Li C. (2006). Uniform colour spaces based on CIECAM02 colour appearance model. *Color Res. Appl.* 31, 320-330.

Maloney, L. T. and Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision* 3, 573-585, <http://journalofvision.org/3/8/5/>, doi:10.1167/3.8.5.

Melgosa, M., Rivas, M. J., Hita, E. and Viénot, F. (2000). Are we able to distinguish color attributes? *Color Res. Appl.* 25, 356-367.

Moreland, J. D. (1972). Peripheral colour vision. In: *Visual Psychophysics, Vol. VII/4 of Handbook of Sensory Physiology* (eds. D. Jameson and L. M. Hurvich), Springer; New York, USA, pp. 517-536.

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

Sharpe, L. T. and Wyszecki, G. (1976). Proximity factor in color-difference evaluations. *J. Opt. Soc. Am.* 66, 40-49.

Sakurai, M., Ayama, M. and Kumagai T. (2003). Color appearance in the entire visual field: color zone map based on the unique hue component. *J. Opt. Soc. Am.* 20, 1997-2009.

Torgerson, W. S. (1958). *Theory and methods of scaling*, Wiley, New York, USA

Witt, K. (2007). CIE colour difference metrics. In: *Colorimetry: Understanding the CIE System* (ed. J. Schanda), Wiley, New York, USA, pp. 83-103.

Xin, J. H., Lam, C. C. and Luo, M. R. (2001). Investigation of parametric effects using medium colour-difference pairs. *Color Res. Appl.* 26, 376-383.

Tables with captions

Table 1. Frequency responses matrix. Experimental condition: size 1 degree, angular distance 30 degrees, colour difference along CIELAB $+a^*$ axis.

		Category label					
Stimulus	ΔC^*	Very Unsatisfied (VU)	Unsatisfied (U)	Lightly Unsatisfied (LU)	Lightly Satisfied (LS)	Satisfied (S)	Very Satisfied (VS)
S1	0	0	0	3	4	19	6
S2	1.17	0	0	3	11	8	10
S3	1.80	0	5	13	8	4	2
S4	2.83	2	9	14	3	3	1
S5	4.35	2	16	11	2	1	0
S6	6.12	6	19	6	0	1	0
S7	7.99	17	13	2	0	0	0
S8	9.25	19	11	2	0	0	0

Table 2. Cumulative proportion matrix. Experimental condition: size 1 degree, angular distance 30 degrees, colour difference along CIELAB $+a^*$ axis.

Stimulus	ΔC^*	Category label					
		Very Unsatisfied (VU)	Unsatisfied (U)	Lightly Unsatisfied (LU)	Lightly Satisfied (LS)	Satisfied (S)	Very Satisfied (VS)
S1	0	0	0	0,09375	0,21875	0,81250	1
S2	1.17	0	0	0,09375	0,43750	0,68750	1
S3	1.80	0	0,15625	0,56250	0,81250	0,93750	1
S4	2.83	0,06250	0,34375	0,78125	0,87500	0,96875	1
S5	4.35	0,06250	0,56250	0,90625	0,96875	1	1
S6	6.12	0,18750	0,78125	0,96875	0,96875	1	1
S7	7.99	0,53125	0,93750	1	1	1	1
S8	9.25	0,59375	0,93750	1	1	1	1

Table 3. Results obtained thanks to the Bonnet's, logistic regression, median and Torgerson-DMT methods. The case presented here is: patches of 1 degree, separated from 30 degrees, on grey background and according to a variation along the axis $+a^*$.

		CATEGORY BOUNDARIES				
METHOD		VU / U	U / LU	LU / LS	LS / S	S / VS
Bonnet's	<i>a</i>	0.38	0.40	0.56	0.62	0.40
	<i>b</i>	-3.19	-1.62	-1.26	-0.66	0.61
	<i>R</i> ²	0.972	0.994	0.834	0.937	0.584
	Category boundaries	8.30	4.09	2.25	1.07	-1.53
logistic regression	<i>b</i> ₁	-5.37	-3.24	-2.39	-1.13	0.90
	<i>b</i> ₂	0.65	0.75	1.16	1.06	0.76
	<i>p</i> value <i>b</i> ₁	0.19	0.13	0.27	0.53	0.64
	<i>p</i> value <i>b</i> ₂	0.23	0.12	0.22	0.28	0.51
	Category boundaries	8.27	4.32	2.06	1.07	-1.18
Median	Category boundaries	7.99	3.96	1.73	1.28	NaN
Torgerson-DMT (with interpolation)	Category boundaries	8.08	4.35	1.68	0.54	-0.89

Table 4. Results obtained using Torgerson-DMT method (in z -values). The case presented here is: patches of 1 degree, separated from 30 degrees, on grey background and according to a variation on the axis $+a^*$.

Sample	z-values	Dispersions
S1	1.00	0.35
S2	0.99	0.46
S3	0.28	0.66
S4	-0.19	0.85
S5	-0.53	0.73
S6	-0.96	0.81
S7	-1.71	0.84
S8	-1.87	0.94
Category boundaries	z-values	
VU / U	-1.64	
U / LU	-0.42	
LU / LS	0.45	
LS / S	0.82	
S / VS	1.29	

Table 5. Results obtained thanks to the logistic regression. The cases presented here is: patches of 1 degree, separated from 30 degrees, on grey background and according to a variation on the axis $-a^*$, $+a^*$, $-b^*$ and $+b^*$. Results are expressed in terms of CIELAB colour differences.

		CATEGORY BOUNDARIES				
Axes		VU / U	U / LU	LU / LS	LS / S	S / VS
$-a^*$	b1	-5.26	-3.33	-3.40	-1.87	0.43
	b2	0.56	0.59	1.19	1.04	0.92
	p value b1	0.19	0.12	0.22	0.33	0.80
	p value b2	0.26	0.13	0.19	0.20	0.44
	Category boundaries	9.40	5.60	2.85	1.81	-0.47
$+a^*$	b1	-5.37	-3.24	-2.39	-1.13	0.90
	b2	0.65	0.75	1.16	1.06	0.76
	p value b1	0.19	0.13	0.27	0.53	0.64
	p value b2	0.23	0.12	0.22	0.28	0.51
	Category boundaries	8.27	4.32	2.06	1.07	-1.18
$-b^*$	b1	-5.28	-4.65	-4.23	-1.80	0.34
	b2	0.42	0.57	0.70	0.59	0.46
	p value b1	0.19	0.15	0.17	0.30	0.83
	p value b2	0.26	0.14	0.14	0.17	0.40
	Category boundaries	12.62	8.21	6.08	3.04	-0.75
$+b^*$	b1	-6.57	-4.36	-2.29	-1.17	1.12
	b2	0.53	0.52	0.47	0.42	0.21
	p value b1	0.22	0.13	0.18	0.40	0.46
	p value b2	0.26	0.14	0.13	0.19	0.54
	Category boundaries	12.33	8.36	4.92	2.77	-5.30

Figures captions

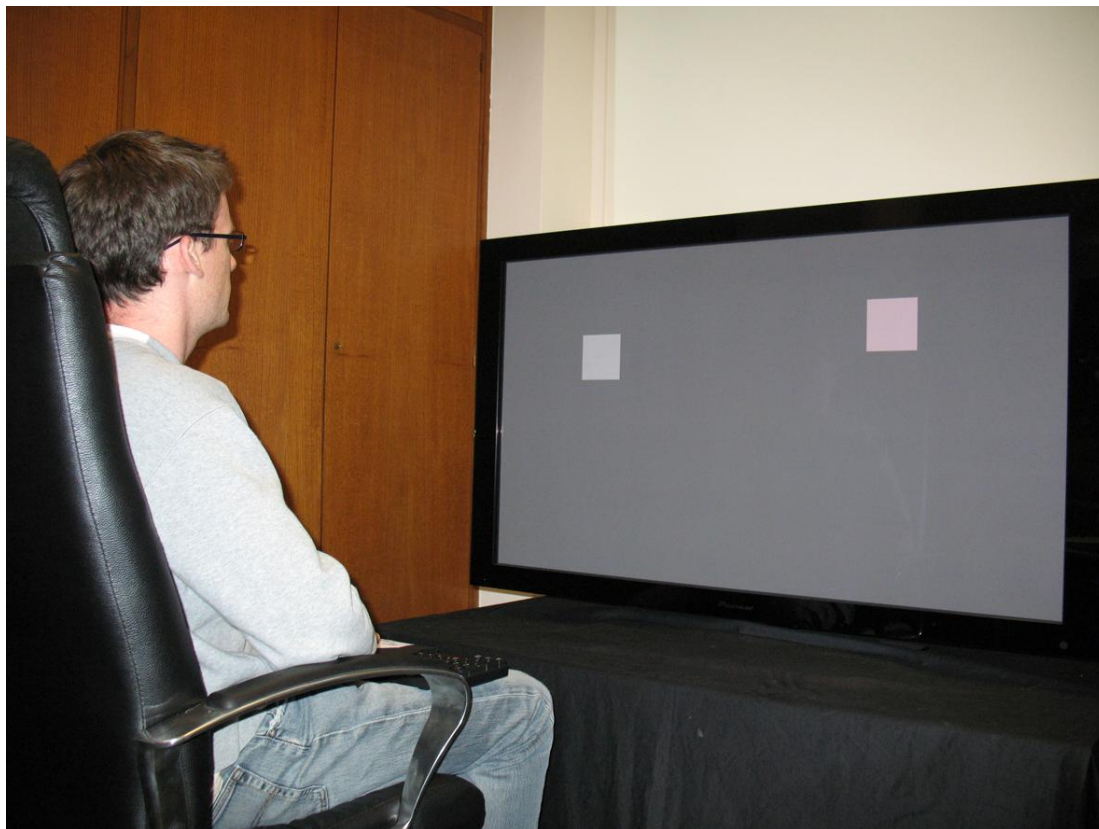


Figure 1: Example of experimental conditions.

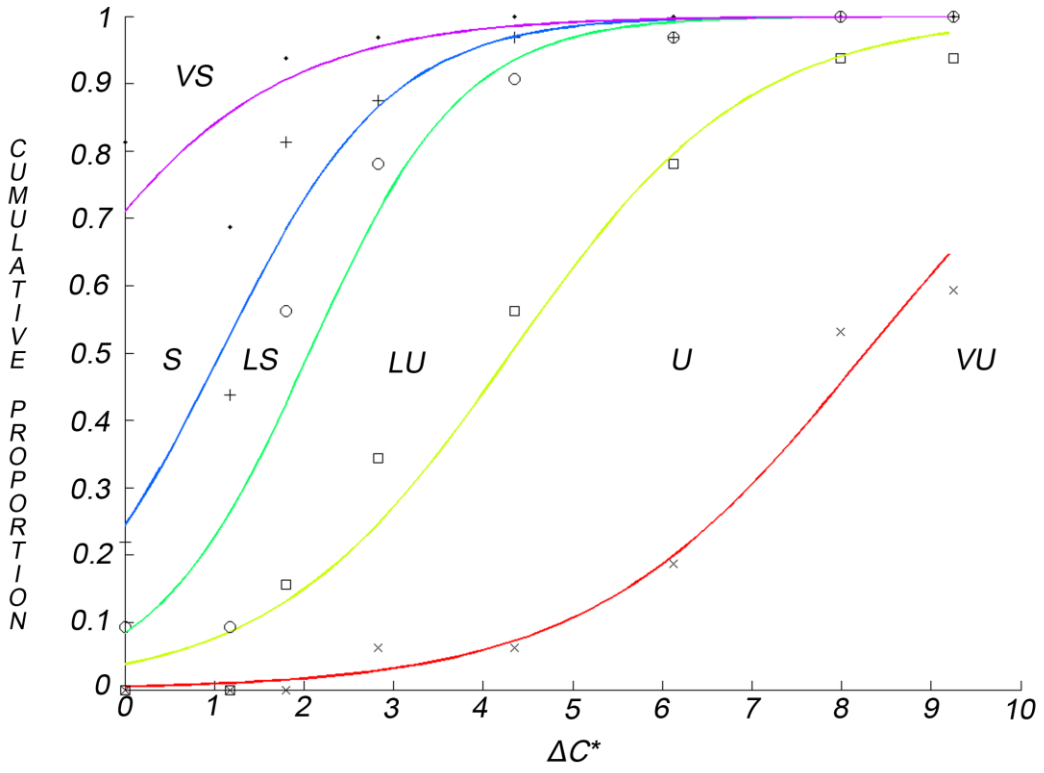


Figure 2: Logistic regression results. Every sigmoid corresponds to a category boundary. Categories are indicated between curves.

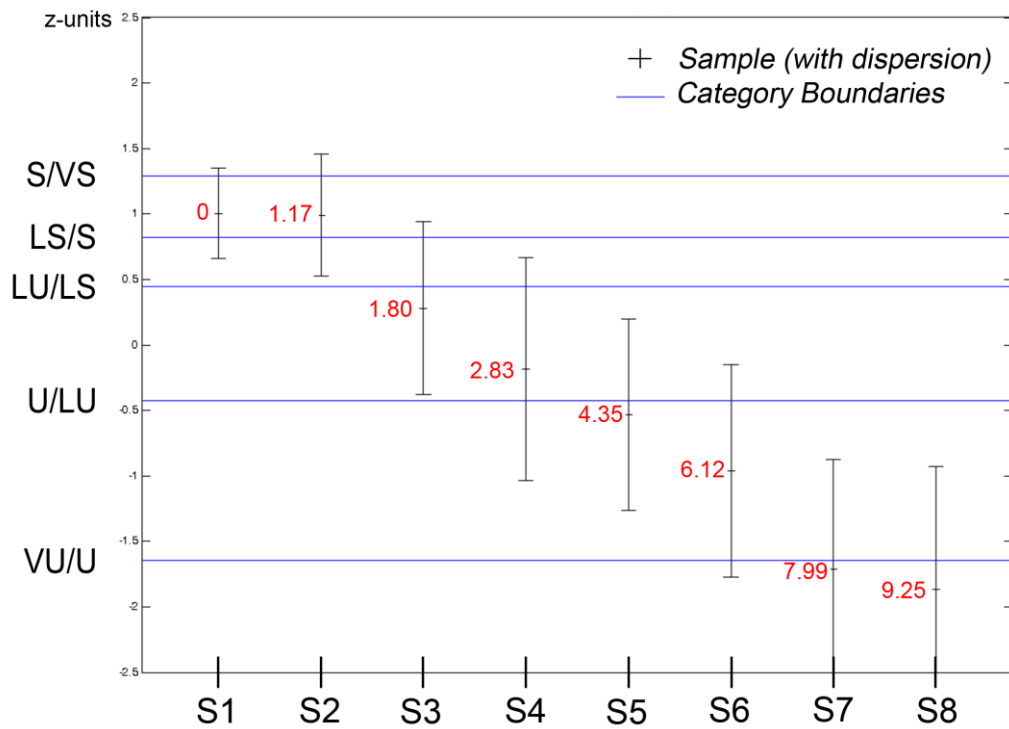


Figure 3: Torgerson-DMT. Results are expressed in z-units. CIELAB ΔC^* magnitude values associated with the test samples are fixed and marked next to the symbols.

Interpolation allows to calculate the ΔC^* magnitude values associated with the category boundaries as given in Table 4.

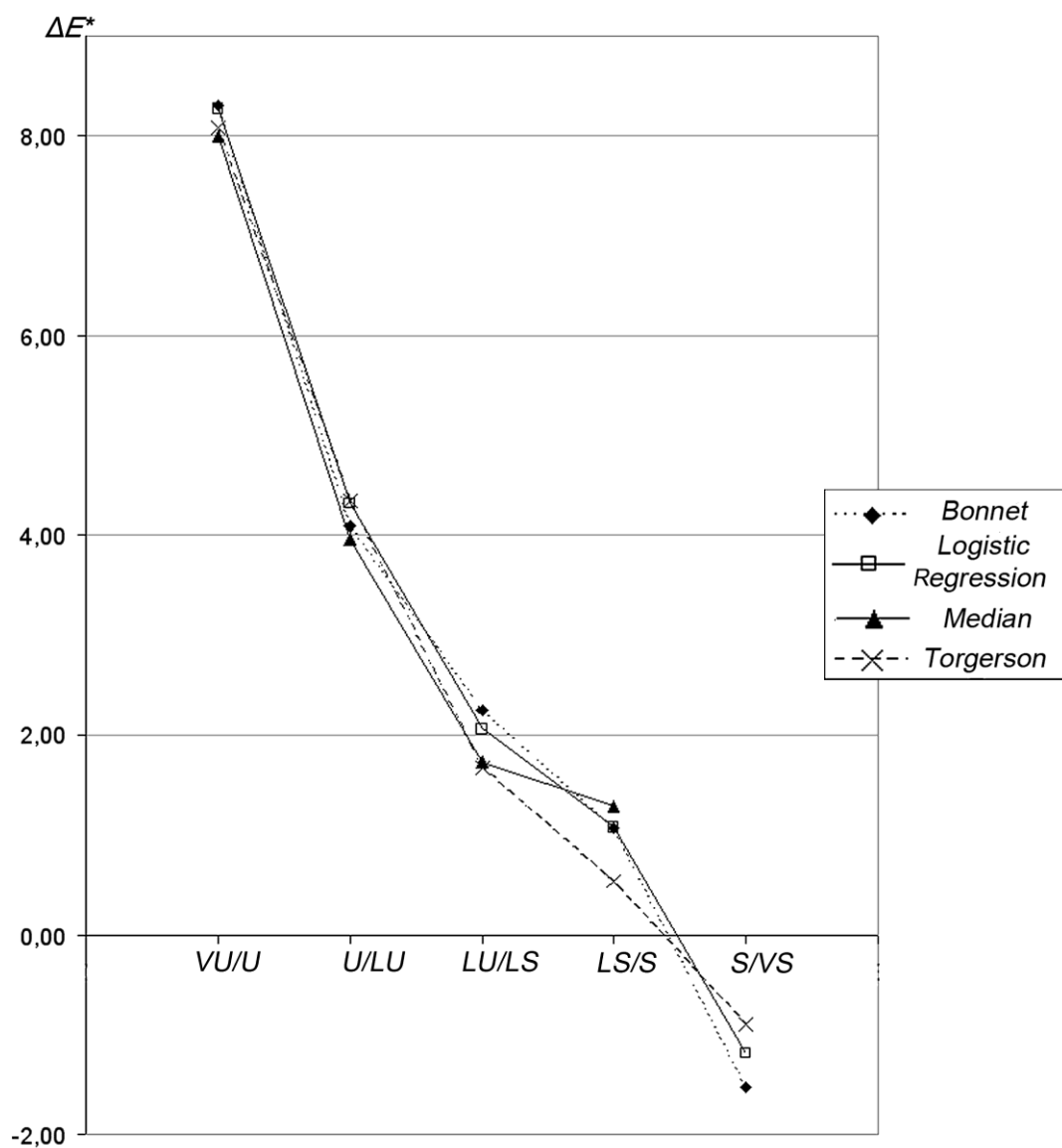


Figure 4: Category boundaries resulting from the Bonnet's, logistic regression, median and Torgerson-DMT methods. The case presented here is: patches of 1 degree, separated from 30 degrees, on grey background and according to a variation along the axis $+a^*$.