An Engineering Model for Color Difference as a Function of Size

Maureen Stone[1], Danielle Albers Szafir[1,2], and Vidya Setlur[1]
[1] Tableau Research, Seattle, WA; [2] Department of Computer Sciences, University of Wisconsin-Madison

Abstract

This work describes a first step towards the creation of an engineering model for the perception of color difference as a function of size. Our approach is to non-uniformly rescale CIELAB using data from crowdsourced experiments, such as those run on Amazon Mechanical Turk. In such experiments, the inevitable variations in viewing conditions reflect the environment many applications must run in. Our goal is to create a useful model for design applications where it is important to make colors distinct, but for which a small set of highly distinct colors is inadequate.

Introduction

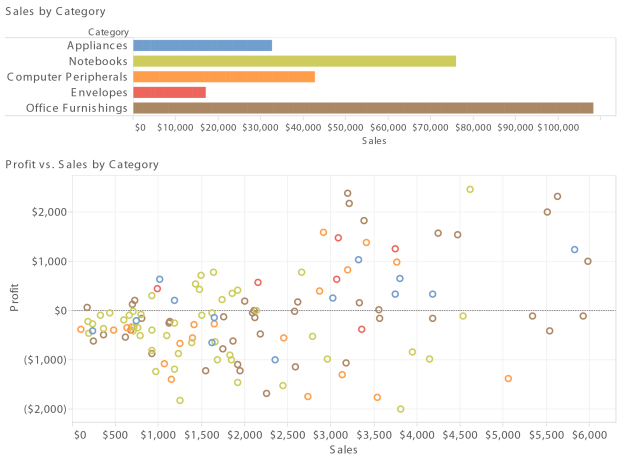

Figure 1. *Colors that are comfortably distinct on bars are more difficult to distinguish on the small scatterplot marks.*

Most color technologies are defined for targets of 2 or 10 degrees [1]. However, designers of color for digital applications have targets of many sizes to consider. While it is well understood that the appearance of color varies significantly with size [5], there are as of yet no practical models to help a practitioner control for this effect. This paper looks specifically at the problem of discriminability, providing a way to estimate how much separation (measured in CIE ΔE units) colors must have to be robustly distinct at different sizes. Our goal is to create a useful model for design applications where it is important to make colors distinct, but for which a small set of highly distinct colors is inadequate.

The ability to distinguish different colorings is especially important in data visualization, where the color indicates a property of the data [10]. For example, in Figure 1, the bar colors, which indicate categories of products, are easy to distinguish. However, for smaller marks such as those in the scatterplots, their component colors become less visibly distinct. In systems like Tableau (`http://www.tableausoftware.com/`), colors are carefully crafted to be robust across sizes. Our goal is to provide better metrics and models for this type of design.

In this work, our goal is to provide a quantitative model of how color discriminability changes as a function of size, with a specific emphasis on discriminability of small perceptual differences, such as just-noticeable differences (JNDs). We base our explorations on a series of crowdsourced experiments, similar to those presented by Szafir *et al.* [11], which explore these phenomena for real users in real viewing conditions. In such experiments, the inevitable variation in viewing conditions reflect the environment many applications must run in. Our goal is to define discriminability in a way that is robust, on average, to these conditions.

This choice represents a direct trade-off: In contrast to other work in this area [3] we are not attempting to model the mechanisms that control how the perceptual system is influenced by color as a function of size. Instead, by measuring this color/size phenomena under more realistic circumstances, we want to derive findings that can be immediately leveraged in practical design.

In the paper, we describe a way to model discriminability as a function of size for target sizes ranging from 6° to $\frac{1}{3}^\circ$ of visual angle. Our noticeable difference function, $ND(p,s)$ is a weighted Euclidean distance in CIELAB space, parameterized by two factors: A threshold p, defined as the percentage of observers who see two colors separated by that value as different, and a size s, specified in degrees of visual angle. A theoretical CIELAB JND, where $p = 50\%$ and $s = 2^\circ$, should correspond to a difference of 1, with equal contributions from L^*, a^* and b^*. For practical design under uncontrolled conditions, we find the required difference, or in our notation, $ND(50,2)$, is closer to 6, with different weightings on L^*, a^* and b^*. As the target size shrinks, the ND value increases and the difference in discriminability along each of the three axis changes unevenly. For 0.33 degrees, the required difference is closer to 11, with an even stronger variation in weightings along the three axes.

Contribution: We empirically evaluate discriminability for 11 different target sizes, ranging from 6° to $\frac{1}{3}^\circ$ of visual angle. Following the model presented by [11], we create a noticeable difference function $ND(p)$ for each size s. We then generalize these results in two ways. First, for a fixed p, estimate $ND(p)$ for an arbitrary size s. This function takes the form $ND(p) = C + K/s$, where C and K are constants obtained from fitting the data for each of our 11 sizes. Second, we generalize this result for arbitrary values of p, creating a general function $ND(p,s)$. The resulting model is also a linear function of inverse size. While further evaluation and refinement is needed, these models provide a simple way to predict discriminability as a function of size.

Related Work

Recent papers by Carter & Silverstein [3, 4] address the problem of discriminability for small colored targets, focusing on those in the range of 120′ to 7.5′ of visual angle. This work leverages reaction time data for a set of identification tasks to understand how the bound of immediate discriminability shifts as a function of size. The resulting formulation communicates a notion of immediate perceptual discriminability, providing parameters for scaling color differences in cone space and for accounting for optical scattering between each small mark and the background as a function of per-cone channel contrast. While we are interested in a larger range of sizes (6° to $\frac{1}{3}^{\circ}$ are discussed in this paper) and more subtle differences, we do incorporate aspects of their model in the design of our experiments.

The sCIELAB work of Zhang & Wandell [12] addresses the problem of evaluating pixel-sized color differences. While an excellent example of a practical model, its focus is pixels in images and does not scale to the range of sizes we are interested in.

That ΔE computed as an Euclidean distance in CIELAB space does not accurately capture color difference is well established. Mahy *et al.*'s evaluation of uniform color differences [8] offers an average value of 2.3 for the JND in CIELAB, in contrast to its theoretical 1.0. Color difference formulations such as CIE94 and CIEDE2000 include parameters to adjust the JND across the color space as a function of hue and chroma [7, 9]. Our work currently assumes uniformity across the space, but this is clearly not true. It will be part of our future work to incorporate some of the insights from these more recent difference formulations, especially the contribution of chroma to our calculations.

Fundamental to our approach is the work by Szafir *et al.* [11], who have demonstrated that evaluating CIELAB based on crowdsourced experiments produces useful results for modeling appearance effects. In their work, they evaluate color differences along the three color axes independently, then rescale CIELAB to create a more robust metric for color difference. We directly follow their procedure for collecting and evaluating color difference judgements of samples jittered along the the L^*, a^*, b^* axes to create a scaled model of CIELAB for each size tested.

Experiment

To rescale CIELAB as a function of size, we require data that measures whether two similar colors appear the same or different. By varying the sizes, colors, and differences, we can calculate scaling factors for the L^*, a^* and b^* axes.

Design

We designed our experiments to use Amazon's Mechanical Turk (`https://www.mturk.com`) infrastructure to crowdsource our experiments. This approach has been validated as being comparable to controlled experiments if sufficient participants are used and care is taken to filter out clearly invalid responses [6, 13, 2]. In addition, creating a model that incorporates the variation in viewing conditions inherent in crowdsourcing is fundamental to our goals.

Participants were shown a series of pairs of colored squares and asked to identify whether the pairs were of the same color or different colors by pressing one of two keys ("f" key if the colors appear the same, and the "j" key if the colors appear different). For each pair, one square was a standard sample, and the second was a "jittered" version of that color, different by a small step along one of the three CIELAB axes. The position of the jittered square was randomized for each stimulus. A set of 52 sample colors were selected by sampling uniformly along the L^*, a^*, and b^* axes. The resulting set is shown in Figure 2. There are 6 L^* steps ranging from 30 to 85. For each L^* value, a^* and b^* were sampled with a spacing of 25; all values that would go out of gamut when jittered were removed. This gave us a basic range of 50 to -50, plus one sample for b^* = -75. While it does not encompass the entire gamut, this set of colors is representative of those used in practical design.

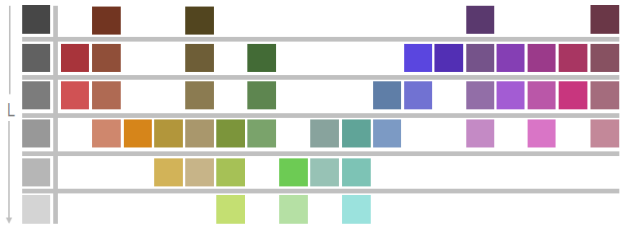

Figure 2. *The 52 sample as distributed in CIELAB space.*

To generate the jittered colors, we defined a jitter step for each size and sampled ±5 steps per axis. This creates 33 color differences per size, including 3 where the color difference was zero. We include zero difference cases both for completeness and to aid in validation. Each participant saw all 33 jitter cases a total of 3 times, but with different colors mapped to each jitter step. We distributed the colors across participants such that we had an equal number of samples for each color × jitter location.

For sizes less than 2 degrees, the jittered distances were modulated based on the Carter & Silverstein recommendations, normalized by their recommended factors such that the 2-degree square step size equaled $1\Delta E$. This helped ensure that we made large enough jitter steps for the small sizes. Step sizes were linearly interpolated for sizes not sampled in the Carter & Silverstein numbers. Optical scattering parameters were not included in this model as we could not uniformly determine whether the difference would result in uniformly positive or negative contrasts agnostic of the standard color. For sizes 2 degrees and larger, a uniform step of $1.25\Delta E$ was used.

We ran our study using a total of 4 experiments, each evaluating three size sets: 0.33, 0.67, and 1 degree; 0.5, 1.25, and 2 degree; 2, 4, and 6 degrees, and 0.4, 0.8, and 1.625 degrees. We replicated the 2 degree value because our initial jitter step for 2 degrees of $1\Delta E$ was found to be too small. In our modeling, we use the results from the larger step. In all cases, the stimuli were presented a fixed distance apart (4 degrees) measured edge to edge. We assumed a standard viewing distance of 24 inches and the HTML default of 96 dpi for pixel resolution (`http://www.w3.org/TR/css3-values/#absolute-lengths`). In most browsers, this will be remapped automatically to compensate for the actual display resolution.

For each experiment, participants first were prompted for their demographic information. Then they were then given a brief tutorial explaining the task at hand. Each participant saw 104 trials, 99 experimental observations and 5 validity trials, in which the sample was presented with a very different color ($\geq 20\Delta E$ difference). There was a 500ms white screen between trials to al-

leviate adaptation effects. As is typical in experiments run on Mechanical Turk, we had to replace roughly 15% of the participants based on our validity criteria, which included correctly identifying the very different cases, the zero difference cases, plus a visual check of the pattern of response. We repeated this process until we had a complete set of observations for our data.

Statistical Analysis

Overall, we analyzed responses from 624 participants (245 female, 339 male, 40 declined to state) between 16 and 66 years of age ($\mu = 33.71$, $\sigma = 11.60$) with self-reported normal or corrected-to-normal vision. Each participant saw each of the 52 stimulus colors twice, with each combination of color difference (jitter amount $\times$ jitter direction $\times$ jittered axis) presented once for each of three sizes. Color $\times$ size $\times$ color difference was counterbalanced between participants. This sampling density will predict discriminability rates for each tested color difference to at worst $\pm 7.5\%$ with 90% confidence.

To verify the validity of our results, we ran an 9-level ANCOVA on the discriminability responses for each sample across all four experiments in the study, treating gender as a covariate, participant id as a random factor to help account for interparticipant variation, and size as a between-subjects factor. We found significant effects of age ($F(1,607) = 8.1342$, $p = .0045$) and question order ($F(1,50826) = 16.7810$, $p < .0001$); however, we saw no systematic variation for either factor. We also saw significant effects of the fixed color's L^* ($F(1,50791) = 1448.323$, $p < .0001$) and b^* ($F(1,50764) = 29.9342$, $p < .0001$) values, but not on the fixed color's a^* value ($F(1,50764) = 0.1621$, $p.6873$); however, only L^* appeared to have a systematic influence on response patterns – discriminability was slightly better for light colors than for dark. Our primary factors – size ($F(10,6741) = 58.2625$,$p < .0001$) and color difference along L^* ($F(1,50756) = 8301.816$, $p < .0001$), a^* ($F(1,50756) = 7819.245$, $p < .0001$), and b^* ($F(1,50756) = 4974.221$, $p < .0001$) — all had a highly significant effect on response.

Predicting Discriminability Thresholds

Based on our data, we can create a parameterized noticeable difference (ND) as a linear function of distance in CIELAB space for each size in our study. Our experiments presented two color patches, a known jitter step apart along either the L^*, a^* or b^* axis, and recorded whether observers said they looked the same or different. We then plotted the jitter step size and the percentage of the responses that indicated it looked "the same." That is, given a distance in CIELAB units between two colors, for each size s, we can predict what percentage of observers p, reported a visible difference. As in the work of [11], we found that a linear model forced through 0 fit this data well. This gives:

$$p = V(s) * \Delta D + e \tag{1}$$

where s is the size, V and D are vector values (L^*, a^* , b^*) and e is experimental and observational error. That is, D is a step in CIELAB space, and V is a vector of three slopes, which are different for L^*, a^*, and b^*. This is shown in Figure 3. Table 1 summarizes the slopes data.

Given this simple model from Equation 1, $ND(p) = p/V$, with ND equivalent to the vector ΔD. For example, to compute the distance vector where 60% of the observers saw a difference,

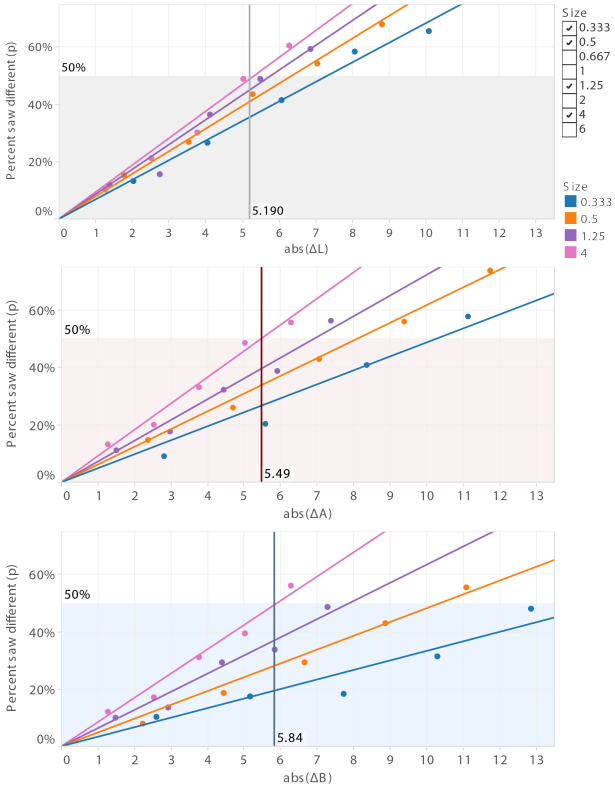

Figure 3. *The slope lines for 4 of the sizes we tested (others removed for legibility). The 50% line is marked, the $ND(50)$ for each of L^*, a^* and b^* axis is the intercept with this line. The $ND(50)$ for the 4-degree stimulus is indicated. All models fit with $p < 0.0001$ except for Δb for size 0.33 ($p = 0.000189$).*

simply divide 0.6 by V, which will return a vector $(\Delta L, \Delta a, \Delta b)$ indicating the steps in LAB space that separate two colors with a 60% reliability. Classically, a *JND* is defined as color difference where 50% of the observers saw a difference, or in our notation, $ND(50)$. Using this formulation and the data in the slopes table, we get the values for $ND(50)$ for each size shown in Table 2. We can use this data to estimate $ND(p,s)$ in two different ways.

Predicting $ND(p,s)$ for a Fixed Value of p

Given a fixed p, we want to predict $ND(p)$ as a function of size. We start by plotting $ND(p)$ for a specific threshold for against size, which shows that the function is non-linear and that the three axes are quite different. For example, Figure 4 shows the plot for $ND(50)$.

Our first result is that a linear fit to 1/size fits this data well, as shown in Figure 5, giving:

$$\text{ND(50,s)} = \text{C(50)} + \text{K(50)/s} \tag{2}$$

Linear regression creates the coefficients for $C(50)$ and $K(50)$ shown in Table 3.

The form of this function makes sense perceptually. As size increases, the K/s term goes to zero, leaving a constant $ND(50)$ of $(5.1, 5.3, 5.3)$. As size decreases below 1, $ND(50)$ increases more rapidly, which matches our observed results.

Changing the value of p we can create the same function but with different coefficients. This provides a two-step model for

	Size (s)										
Axis	0.333	0.4	0.5	0.667	0.8	1	1.25	1.625	2	4	6
L^*	0.068	0.069	0.078	0.081	0.090	0.083	0.089	0.085	0.100	0.096	0.090
a^*	0.051	0.054	0.062	0.067	0.064	0.073	0.073	0.072	0.085	0.091	0.097
b^*	0.034	0.042	0.050	0.051	0.055	0.061	0.064	0.066	0.073	0.086	0.086

Table 1. *$V(s)$ for each size and axis*

	Size (s)										
Axis	0.333	0.4	0.5	0.667	0.8	1	1.25	1.625	2	4	6
L^*	7.321	7.267	6.435	6.180	5.531	6.017	5.643	5.903	5.010	5.187	5.574
a^*	9.901	9.268	8.052	7.429	7.837	6.897	6.821	6.906	5.917	5.488	5.149
b^*	14.837	12.019	10.101	9.747	9.091	8.197	7.764	7.587	6.831	5.841	5.834

Table 2. *ND for $p = 50\%$ for each size and axis*

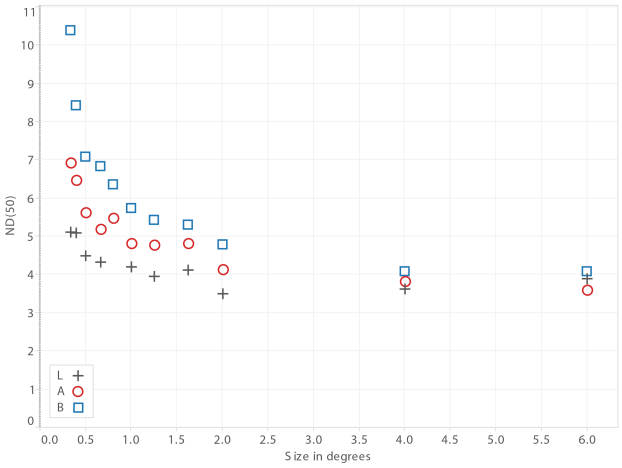

Figure 4. *$ND(50)$ plotted against size for each of our tested sizes for each axis. L^* is gray plus, a^* is red circle, b^* is blue square.*

Axis	$C(50)$	$K(50)$
L^*	5.079	0.751
a^*	5.339	1.541
b^*	5.349	2.871

Table 3. *C and K coefficients for $ND(50)$*

discriminability as a function of size. First, compute $ND(p)$ for the desired p, then use linear regression to define the coefficients for the following equation:

$$ND(p,s) = C(p) + K(p)/s \tag{3}$$

Generalizing the Model

In the previous section, we created a model of $ND(p,s)$ for a fixed p. Here we generalize this model so we can predict $ND(p,s)$ for an arbitrary p without having to calculate and fit the ND(p) data. To do this, we need to predict the slopes shown in Figure 3 and Table 1 from the size. Based on the results in the previous sections, we would expect to see a solution in the following form:

$$V(s) = p/ND(p) = p/(C(p) + K(p)/s) \tag{4}$$

where $C(p)$ and $K(p)$ are the coefficients in Equation 3.

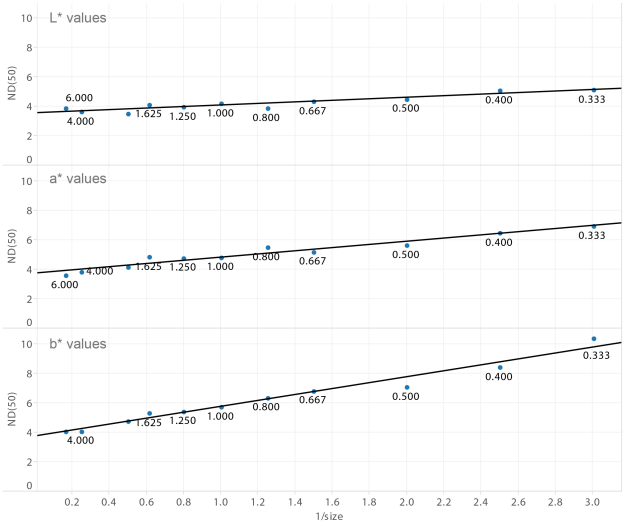

Figure 5. *The plot of $ND(50)$ for each of the 11 sizes vs. $1/size$ for each of L^*, a^* and b^*. ($R^2_L = .849696, p_L < 0.0001$; $R^2_a = .942234, p_L < 0.0001$; $R^2_b = .970395, p_b < 0.0001$)*

Plotting slope, V, as a function of size gives the non-linear distribution shown in Figure 6.

Using linear regression to fit $1/V$ as a function of $1/s$ gives the results shown in Figure 7, or:

$$1/V(s) = A + B/s \tag{5}$$

This gives a general specification for $ND(p,s)$:

$$ND(p,s) = p(A + B/s) \tag{6}$$

where s is size in degrees, p is in the range 0 to 1, and the values for A and B are shown in Table 4.

Axis	A	B
L^*	10.16	1.50
a^*	10.68	3.08
b^*	10.70	5.74

Table 4. *A and B coefficients for Equation 5*

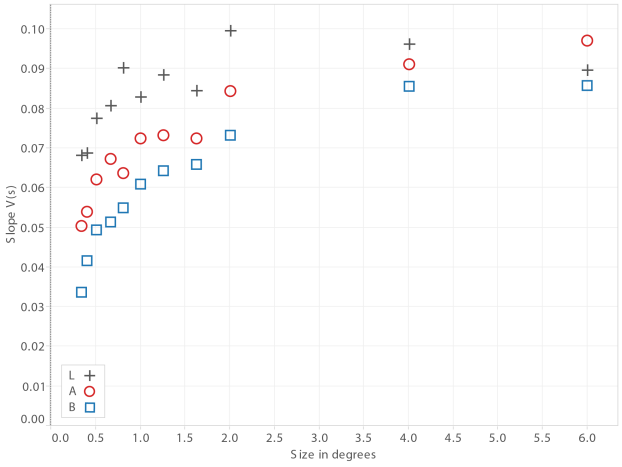

Figure 6. *The distribution of the slope, V vs. size for our data. Gray cross is L^*, red circle is a^*, blue square is b^*.*

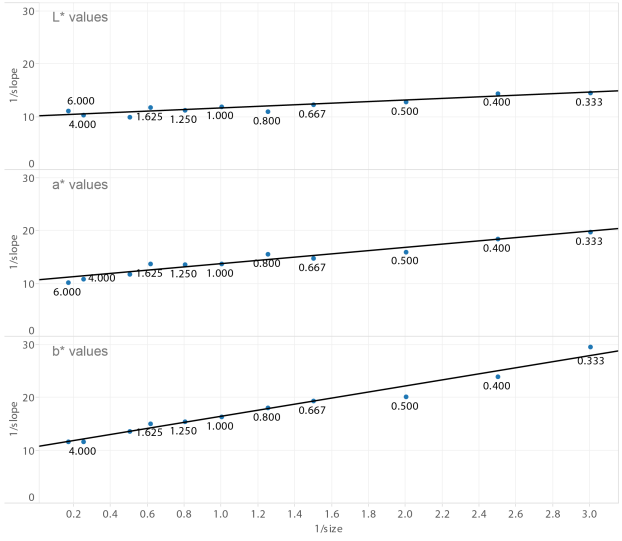

Figure 7. *Linear fit to $1/V$ vs $1/size$ for each of L^*, a^* and b^*. ($R^2_L = .849696, p_L < 0.0001$; $R^2_a = .942234, p_L < 0.0001$; $R^2_b = .970395, p_b < 0.0001$).*

Discussion

To visualize this model, we have used Tableau Software's visual analysis system (`http://www.tableausoftware.com`). The slopes for each size, $V(s)$ were computed from the experimental data, then used to create an interactive model in Tableau. We defined a function, $ND(p,s)$, along with a variable parameter for p. By changing p, we can see the different regression lines modeling 3 for different fixed values of p. Figure 8 shows the different $ND(p,s)$ regression lines for $p = 50$ for each axes. The shaded bands show the variation in ΔL^*, Δa^* and Δb^* over the range of sizes, with the band size increasing for L^* vs. a^* vs. b^*.

Increasing p moves the regression lines up and decreasing it moves them down. For example, setting $p = 35$ (Fig. 9), we see that smaller delta L^*, a^* and b^* values are needed to guarantee 35% discriminability. There remains a good linear fit to these new points, but the predicted ND values are lower. In this figure,

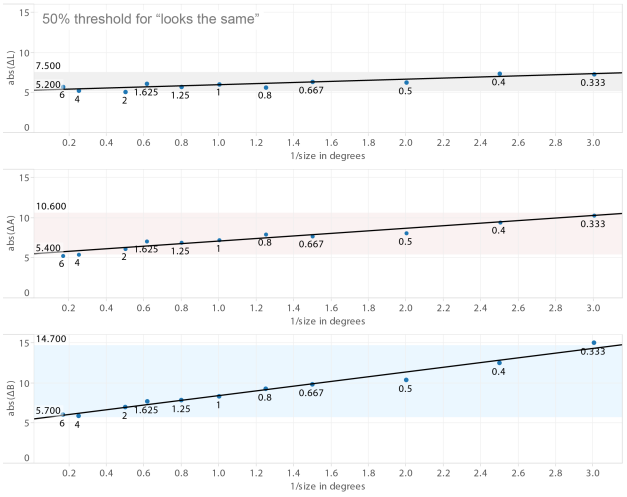

Figure 8. *The figure shows the color difference step needed for 50% discriminability $(ND(50))$ for each axis as a linear model of $1/size$. Colored bands are labeled with the range of color difference values for each axis.*

the bands continue to encode the 50% range, for reference.

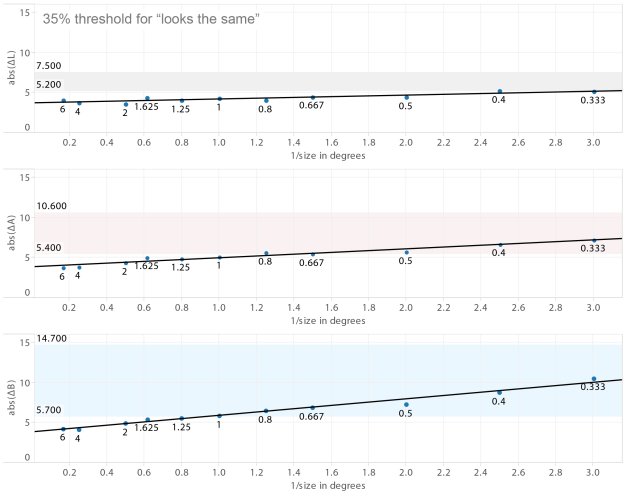

Figure 9. *Same as Figure 8, but with $p = 35$. The reference bands still show the ranges for $p = 50$, for comparision*

These results can also be visualized by plotting sample color patches and observing the difference. The challenge with this is that not only discriminability but overall appearance changes as colors get small—small stimuli appear less colorful. In Figure 10, both the large and small patches are stepped according to the parameters of our algorithm. Ideally, the color differences will seem the same independent of size. For comparison, the small patches are also shown with the same color steps as the large patches, and should appear less different.

While we used the same basic procedure described in [11] for our experiments, we did not get the same ND(50) values for our 2-degree patches as they did for theirs. They estimated $ND(50) = (4, 5.3, 5.8)$ and our results are $(5, 5.9, 6.8)$ for similar populations (Amazon Turk workers). We have started to explore this difference. We first hypothesized that combining three

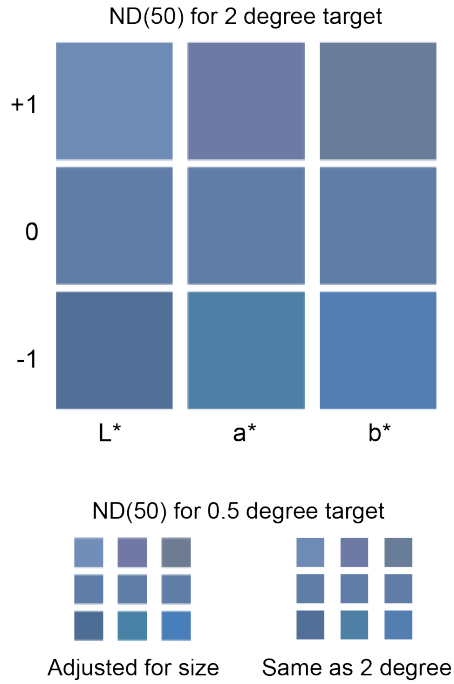

Figure 10. *Assuming a viewing distance of 18 inches, the large patches are 2 degree squares and the small ones 0.5 degrees. Vertically adjacent patches are one $ND(50)$ step different as computed from our formulas. For comparision, the 0.5 degree squares are also drawn with the 2-degree values.*

sizes in one experiment may have made the comparisons more difficult to perform. By repeating our study with only one size, however, we did not find a significant difference from our previous result. Exploring further, we discovered that the previous work more carefully trained participants to look for smaller color differences. Therefore, it may simply be that our data includes more people who identify only the largest differences. Since this is consistent across all our experiments, this doesn't invalidate our models. However, the $ND(p,s)$ values reported here may be larger than necessary.

Conclusion and Future Work

The work presented in this paper offers a simple model for computing color difference as a function of size. While the results are preliminary, this sort of data-driven modeling shows strong promise for creating practical results. Our data indicates that a minimum step in CIELAB of between 5 and 6 is what is needed to make two colors visibly different for large shapes (2-degree or larger), which matches well with the intuitions developed through design practice by the authors. That there is an asymmetry between L^* and the two axes defining colorfulness (a^* and b^*) also matches our experience; small shapes need to be much more colorful to be usefully distinct.

Future work will include studies to refine the model parameters, and to explore the effect of background color. In parallel, we intend to put these results into practice, using them as design and evaluation guidelines for color sets used in visualization. This will help us best understand what parts of the model need further refinement.

Acknowledgments

We would like to thank Justin Talbot for his help with the data modeling and analysis.

References

[1] R.S. Berns. *Bilmeyer and Saltzman's Principles of Color Technology, third edition*. Wiley-Interscience publication. Wiley, 2000.

[2] M. Buhrmester, T. Kwang, and S.D. Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[3] Robert C. Carter and Louis D. Silverstein. Size matters: Improved color-difference estimation for small visual targets. *Journal of the Society for Information Display*, 18(1):17, 2010.

[4] Robert C Carter and Louis D Silverstein. Perceiving color across scale: great and small, discrete and continuous. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 29(7):1346–55, July 2012.

[5] Mark D Fairchild. *Color appearance models*. John Wiley & Sons, 2013.

[6] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 203–212. ACM, 2010.

[7] M.R. Luo, G. Cui, and B. Rigg. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Res. Appl.*, 26(5):340–350, 2001.

[8] M. Mahy, L. Van Eycken, and A. Oosterlinck. Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV. *Color Research & Application*, 19(2):105–121, 1994.

[9] AR Robertson. Historical development of CIE recommended color difference equations. *Color Res. Appl.*, 15(3):167–170, 2007.

[10] Maureen Stone. In color perception, size matters. *IEEE Computer Graphics & Applications*, 32(2):8–13, March 2012.

[11] D.A. Szafir, M. C. Stone, and M. Gleicher. Adapting color difference for design. In *Proceedings of the 22th Color and Imaging Conference*. Imaging Sciences and Technology, November 2014.

[12] Xuemei Zhang and Brian A Wandell. A spatial extension of cielab for digital color-image reproduction. *Journal of the Society for Information Display*, 5(1):61–63, 1997.

[13] S. Zuffi, C. Brambilla, G.B. Beretta, and P. Scala. Understanding the readability of colored text by crowd-sourcing on the web. Technical report, External HPL-2009-182, HP Laboratories, 2009.

Author Biographies

Maureen Stone is a Research Scientist at Tableau Software, where her work focuses on enhancing the effectiveness of information visualization by blending principles from perception and design. Prior to joining Tableau, she worked first at Xerox PARC, then independently as StoneSoup Consulting. She has a BS and MS in Computer Engineering from the University of Illinois, and a MS in Computer Science from Caltech.

Danielle Albers Szafir received her BS in Computer Science from the University of Washington (2009) and is currently completing her PhD in Computer Sciences at the University of Wisconsin-Madison under Prof. Michael Gleicher. Her work focuses on understanding scalability and interpretability in visualization design.

Vidya Setlur is a research scientist at Tableau Software. Previously, she worked as a principal research scientist at the Nokia Research Center for 7 years. She earned her doctorate in Computer Graphics in 2005 at Northwestern University. Her research interests lie at the intersection of natural language processing (NLP) and computer graphics, particularly in the area of data semantics, iconography and content retargeting.