An Engineering Model for Color Discriminability as a Function of Size

Abstract

This work describes a first step towards the creation of an engineering model for color discriminability as a function of size. Our approach is to non-uniformly scale CIELAB using data from crowd-sourced experiments, such as those run on Amazon Turk. In such experiments, the inevitable variations in viewing conditions reflect the environment many applications must run in. Our intent is to define discriminability in a way that is robust, on average, to these conditions. We make no claim that our model describes color perception with a high degree of precision. 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

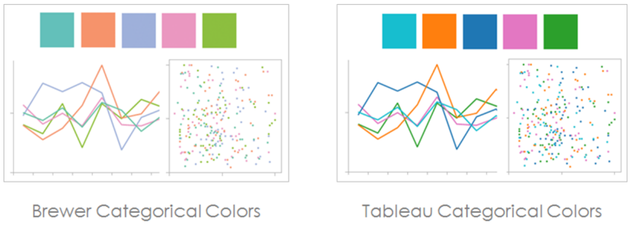

Figure 1: Problems with color discriminability at smaller sizes. Here, both Brewer and Tableau swatch sets are easily discriminable at large sizes. However, as the marks grow smaller, their component colors become increasingly difficult to distinguish.

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 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.

As the size of colored targets becomes smaller, the ability to distinguish between the color of these targets degrades. This is especially problematic in fields such as visualization, where the ability to interpret an image is heavily based on our abilities to distinguish between marks of different colors and sizes [8]. For example, in Figure 1, both swatch sets are easily discriminable at large sizes. However, as the marks grow smaller, their component colors become increasingly difficult to distinguish. In systems like Tableau (`http://www.tableausoftware.com/`), designers have made a conscious effort to address this issue. However, these colors and adjustments were carefully crafted, relying heavily on extensive expertise.

In this work, our goal is to provide a quantitative model of how color discriminability shifts as a function of size, with a specific emphasis on how discriminability functions at small perceptual differences, such as just-noticeable differences (JNDs). We base our explorations on a series of crowdsourced experiments, similar to those presented by [anonymous], focusing on gauging 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 hope 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 in the range 6 degrees to $\frac{1}{3}$ degree of visual angle. Our noticeable difference function, $ND(p,s)$ is a weighted Euclidean distance in CIELAB space, parameterized to a threshold p, defined as the percentage of observers who see two colors separated by that value as different, and by a size s, specified in degrees of visual angle. For a JND, this threshold would be 50%, and as CIELAB was specified for 2-degree targets, this distance would be 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 5.3, with slightly 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, but the weights on the three axes are very different.

Contribution

We performed a set of experiments to evaluate discriminability for 11 different target sizes, ranging from 6 degrees to 120‘ of the visual angle. Using techniques of [anonymous], we create a noticeable difference function $ND(p)$ for each size s. We then generalize these results in two ways:

- For a fixed p, estimate $ND(p)$ for an arbitrary size s. This function takes the form $ND_p(s) = C + K/s$, where K and C are constants obtained from fitting the data for each of our 11 sizes.
- For a more general formulation, we first estimate $ND(p)$ for a specific size, then use this to compute $ND(p,s)$.

Early evaluation of these results indicates that, while more work is needed, they can be useful for determining discriminability as a function of size.

Related Work

Recent papers by Carter and Silverstein [3, 4] address the problem of discriminability of small colored targets, focusing on those in the range of 120 to 7.5 minutes of the visual angle. This work leverages reaction time data for a set 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. We are interested in a larger range of sizes (6° to $\frac{1}{3}^\circ$ are discussed in this paper), and more subtle differences. However, we do incorporate aspects of their model in the design of our experiments.

The sCIELAB work of Zhang and Wandell [9] 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 [7] 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. 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 [anonymous], who have demonstrated that rescaling CIELAB based on crowd sourced experiments produces useful results. 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 colors and the 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 equivalent to controlled experiments if sufficient participants are used and care is taken to filter out clearly invalid responses [6, 10, 2]. In addition, creating a model that is robust to the variation in viewing conditions inherent in crowdsourcing is fundamental to our goals.

Participants were shown a series of pairs of squares and asked to identify whether the pairs were of the same color or different colors by pressing one of two keys. For each pair, one square was a standard sample, and the color of the second "jittered" square was adjusted by a fixed amount from the color of the first along a given axis. The position of the jittered square was randomized for each stimulus. The set of 52 sample colors were selected from a uniform distribution in LCh, reduced to avoid going out of gamut when jittered. The resulting set is shown in Figure 2, selected from 6 lightness steps, 18 hue steps and and 3 chroma steps: 0, 25, 50.

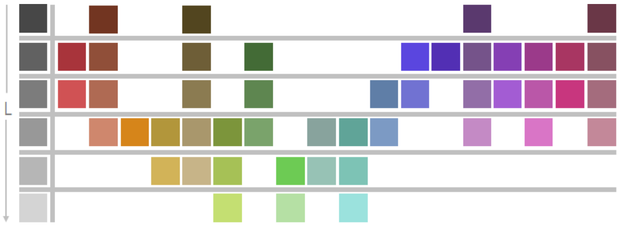

Figure 2: The resulting set of sample colors from a uniform distribution in LCh, selected from 6 lightness steps, 18 hue steps and and 3 chroma steps: 0, 25, 50.

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 was too small. In our modeling, we use the results from the second experiment. In all cases, the stimuli were a fixed distance apart, measured edge to edge. Participants each saw a subset of these colors.

The jittering was done uniformly per axis, with each color adjusted by integer steps, up to ±5 steps per axis. For sizes less than 2 degrees, the jittered distances were modulated based on the Carter & Silverstein recommendations, normalized such that the 2-degree square step size equaled $1\Delta E$ for all squares less than 2 degrees wide. This helped ensure that we made large enough jitter steps. 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 that was agnostic of the constant color. For sizes greater than 2 degrees, a uniform $1.25\Delta E$ step was used.

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 of drastically different RGB color (20 or more ΔE difference). There was a 500ms white screen between trials to alleviate 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. We repeated this process until we had a complete set of observations for our data.

Method

For each experiment, 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 × jitter direction × jittered axis) presented once for each of three sizes. Color × size × color difference was counterbalanced between participants. This sampling density will predict discriminability rates for each tested color difference to at worst ±7.5% with 90% confidence.

To verify the validity of our results, we ran an 9-level AN-

COVA on the discriminability responses for each sample across all four experiments in the study, treating gender as a covariate to 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

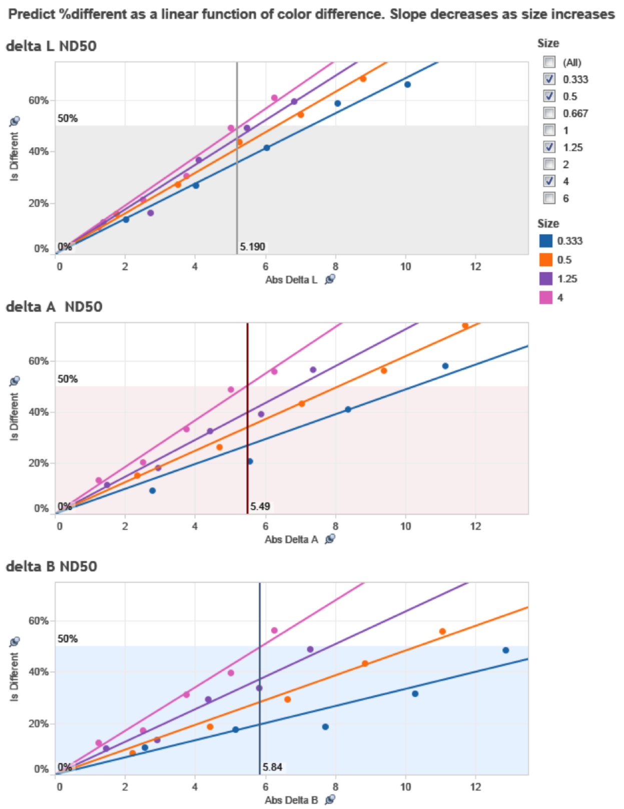

Figure 3: The slope lines for 4 of the sizes we tested (others removed for legibility). The 50% line is marked, the ND_{50} for each of L^*, a^* and b^* axis is the intercept with this line. The ND_{50} for the 4-degree stimulus is indicated.

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. This gives:

$$p = V(s) * \Delta C + e \qquad (1)$$

where V and C are vector functions and e is experimental and observational error. As in the work of [anonymous], we found a linear model forced through 0 fit this data well. That is, C 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. All values fit with $p < 0.0001$ except for Δb for size 0.33 ($p = 0.000189$).

Given this simple model from Equation 1, $ND(p) = p/V$, with ND equivalent to the vector ΔC. For example, to compute the distance vector where 60% of the observers saw a difference, 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. For a fixed p, we write ND_p, so for a JND, we would use ND_{50}. Now we use this data to create a model for $ND_p(s)$, which is ND for p% reliability at a given size s. Based on this model, we get the following values for ND_{50} for each size as shown in Table 2.

We now use this data to estimate $ND(p,s)$ in two different ways.

Predicting ND for a particular threshold as a function of size

Given a fixed p, we want to predict $ND_p(s)$, where the subscript p indicates a specific ND value for a fixed p, rather than as a function of p. We achieve this by plotting ND_p for different sizes which is a non-linear function of size as shown in Figure 4 for ND_{50}.

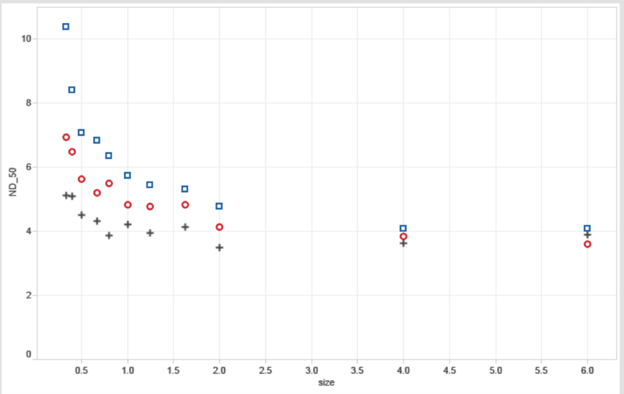

Figure 4: ND_{50} plotted against size for each of our tested sizes for each axis (L^* is gray plus, a^* is red circle, b^* is blue square. Note that for smaller sizes, the three axis are quite different. The non-linearity as size decreases is much more significant in a^* and b^* than for L^*.

We find a function of 1/size gives us a good fit (Figure 5). Linear regression creates the coefficients shown in Table 3, all of which provide a significant fit to the data ($R_L^2 = .849696$,$p_L < 0.0001$; $R_a^2 = .942234$,$p_L < 0.0001$; $R_b^2 = .970395$,$p_b < 0.0001$). That is, $ND_{50}(s) = C_{50} + K_{50}/s$ (Remember that this is a vector equation in L^*, a^* and b^*).

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

	Size										
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 2: ND for $p = 50\%$ for each size and axis

	Size										
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

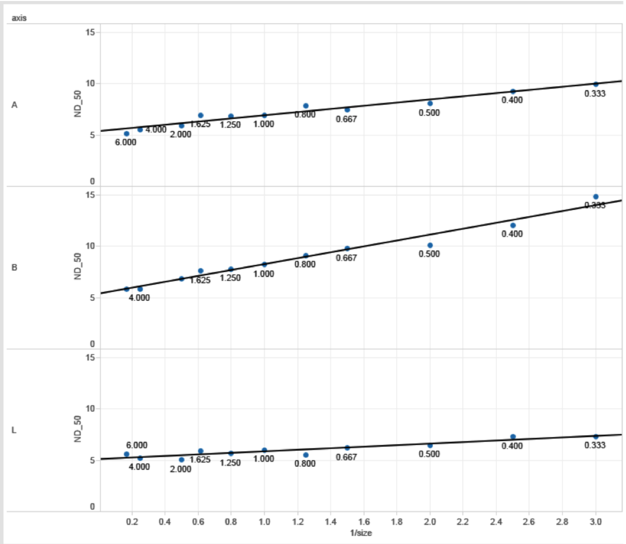

Figure 5: The plot of ND_{50} for each of the 11 sizes vs. $1/size$.

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

Table 3: C and K coefficients for ND_{50}

Axis	C_{50}	K_{50}
L^*	5.07857	0.751199
a^*	5.33884	1.54136
b^*	5.34949	2.87127

For different values of p, we also get good fits, but with different coefficients. This provides a two-step model for discriminability as a function of size. First, compute $ND(p)$ for the desired p, then fit this value with an equation of the form:

$$ND_p(s) = C_p + K_p/s. \tag{2}$$

Generalizing the Model

In the previous section, we created a model of $ND(s)$ for a fixed p. Now let us formulate $ND(p,s)$. Given $ND(p) = p/V(s)$, we need to be able to estimate $V(s)$ for an arbitrary size. Based on the results in the previous section, we would expect to see

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

where $C(p)$ and $K(p)$ are the coefficients in Equation 3, specified here as functions rather than constants. Figure 6 is a plot of V vs. size.

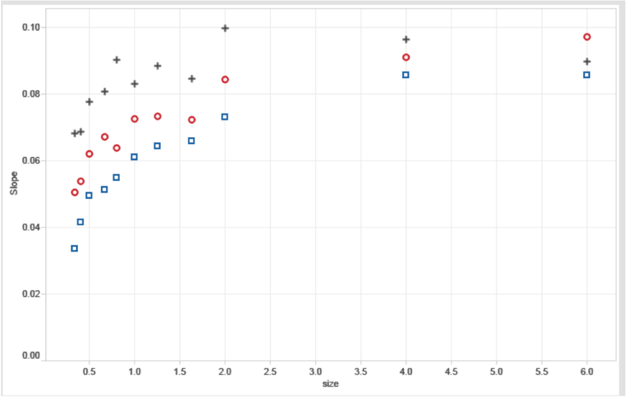

Figure 6: The distribution of V(s) vs. size for our data. Gray cross is L^*, red circle is a^*, blue square is b^*.

Based on Equation 3 we look for a fit to $1/V$ as a function of $1/s$. That is,

$$1/V(s) = A + B/s \tag{4}$$

The resulting fit is shown in Figure 7, with the coefficients show in Table 4. This model also characterizes our data well, providing an equivalent fit quality to the ND_{50} fits ($R_L^2 = .849696, p_L < 0.0001$; $R_a^2 = .942234, p_L < 0.0001$; $R_b^2 = .970395, p_b < 0.0001$).

Table 4: A and B coefficients for Equation 4

Axis	A	B
L^*	10.1571	1.5024
a^*	10.6777	3.08273
b^*	10.699	5.74253

This gives us a simple specification for the coefficients in Equation 3, with $C(p) = pA$ and $K(p) = pB$.

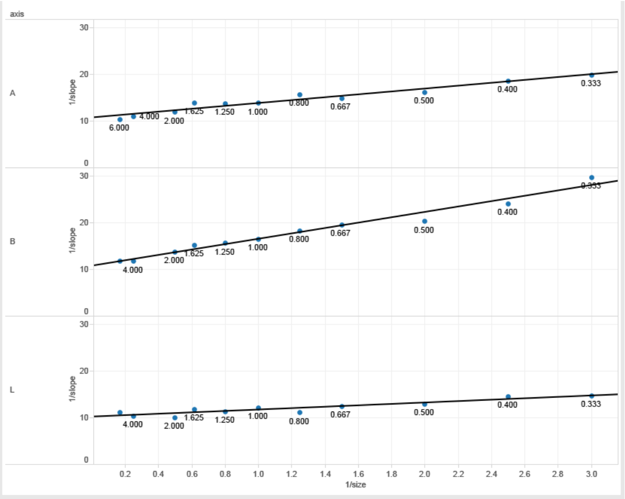

Figure 7: Linear fit to $1/V(s)$ vs. $1/size$

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 independently and read into this workbook as data. We then defined a function, $ND(p)$, and along with a variable parameter for p. Figure 8 shows the different $ND_p(s)$ lines for $p = 50$. The shaded band show the variation in delta L^*, a^* and b^* over the range of sizes. Notice how much wider the band is for b^* vs. a^* vs. L^*.

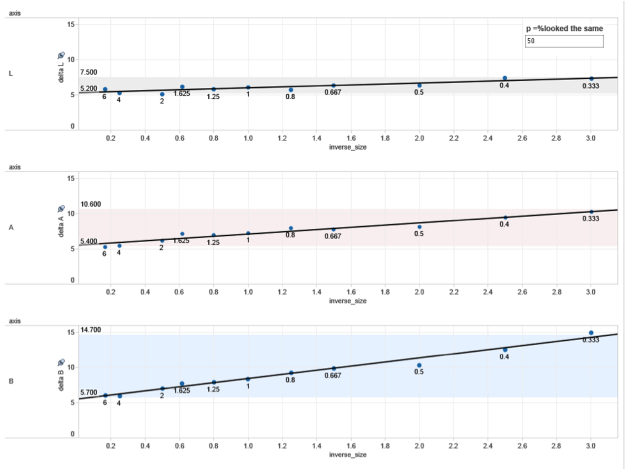

Figure 8: The figure shows the delta value needed for 50% discriminability (ND_{50}) for each axis as a linear model of $1/size$. Colored bands are labeled with the range of delta values for each axis.

By adjusting p, we see that the lines move up and down, but the shape remains similar. In Figure 9, we have set $p = 35$. Now smaller delta L^*, a^* and b^* values are needed to guarantee 35% discriminability. There remains a good linear fit to these new points, as would be expected. (The bands still show the 50% range, for comparison)

Another way to visualize these results is to plot color patches and observe the difference. One challenge with this is that not only discriminability but overall appearance changes as colors get small. Small stimuli generally appear less colorful. However, Figure 10 is an attempt to illustrate our results as colors. Both the large and small patches are stepped according to the parameters of our algorithm. The question is whether the differences seem the same independent of size? For comparison, the small patches are also shown with the same color steps as the large patches.

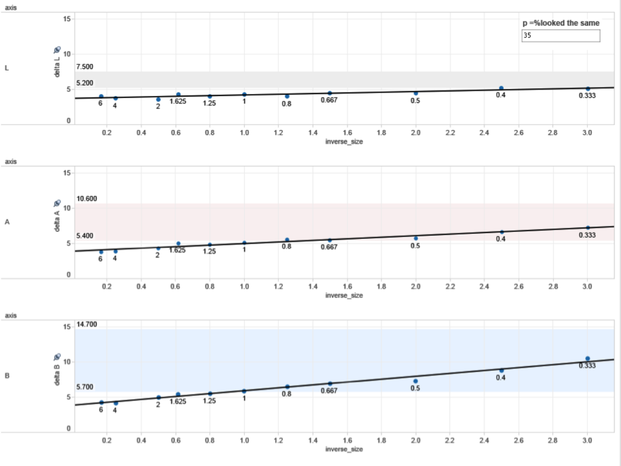

Figure 9: Same as Figure 8, but with $p = 35$.

While we applied the procedures recommended by [anonymous] for scaling CIELAB, we did not get the same results 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 can only speculate about what caused the difference. One possibility is the difficulty of the task. Our experiment used smaller jitter steps, and therefore more comparisons looked "the same." People may have therefore decided the best strategy was to only mark as different colors that were very different. We also found that having a range of sizes made the task harder, and that which sizes were combined had some effect on the results. A better approach might have been to use a constant size and more experiments with fewer trials, which would have been more similar to the [anonymous] study.

Conclusion and Future Work

The work presented in this paper offers a simple model for adjusting color discriminability as a function of size. While the results are preliminary, this sort of data-driven modeling shows strong promise for creating practical results.

This work can be improved by further data collection and analysis, but our results are simple and promising enough to also try out in practice. Our data indicates that a minimum step in CIELAB of between 5 and 6 is what is needed to make two colors visibly different, which matches well with the intuitions developed through design practice by the authors. That this increases somewhat for L^* as the target shrinks is also expected. However, changes along the dimensions controlling colorfulness (a^* and b^*) must be much larger, matching our experience that small targets need to be much more colorful to be usefully distinct.

For design purposes, it is much easier to use LCh rather than the CIELAB axis. But the asymmetric scaling of a^* and b^* will introduces changes in the hue vector with respect to the original CIELAB specification. Either this will be a feature in that it accurately models perception, or it will cause hue shifts that will need

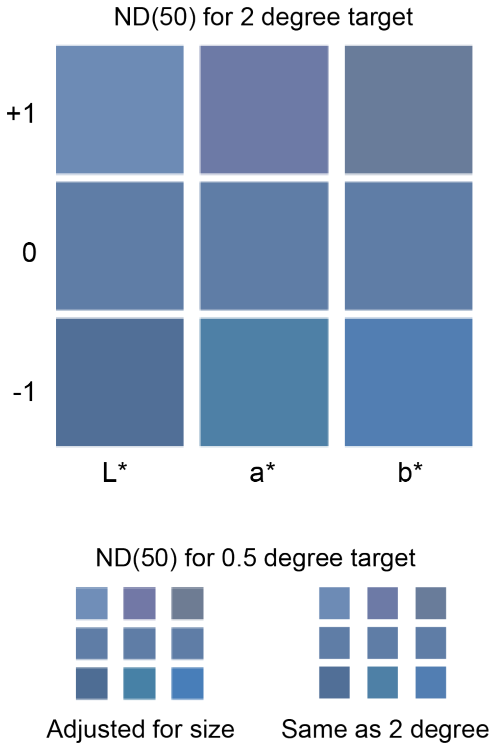

Figure 10: The large patches are intended to be 2 degree squares, with one ND_{50} step in each direction as computed from our formulas. The small patches are 0.5 degrees, and the corresponding ND_{50} steps are larger. For comparison, the same color steps are also shown on the small patches.

compensation. We have already started exploring looking at our results in this context, which will let us more easily apply them to some of our real-world design problems. This will help us best understand what parts of the model need further refinement.

References

[1] RS Berns. Billmeyer and saltzmans. *Principles of Color Technology*, 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. 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.

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

[9] 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.

[10] 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, August 6 2009, http://www. hpl. hp. com/techreports/2009/HPL-2009-182. html, 2009.