Supplemental Material:
Color Compatibility From Large Datasets

Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann

Project URL: `www.dgp.toronto.edu/~donovan/color/`

1 Unmixing color preferences

In the paper, we plot the average ratings of all themes containing each color. However, this mixes together the contributions of each color to the rating. Here we consider an approach to "unmixing" the effect of color preferences on theme ratings.

We discretize hues, and treat each distinct hue j as having a hidden "quality" q_j. Suppose a theme $\mathbf{t}$ has rating r. We model this theme's rating as arising from the average of the qualities of the $N \leq 5$ colors of a theme as:

$$r = \sum_{j \in \mathbf{t}} q_j / N \tag{1}$$

The Kuler data provides us with a large collection of pairs of themes and rankings. Each theme has a rating and set of colors, yielding a linear equation of the form of Eqn. 1. We can directly estimate the qualities q of each color by solving the resulting system of equations in a least-squares sense. Only saturated and light colors are considered ($c_{sat} > \tau_{sat}$ and $c_{val} > \tau_{val}$), and themes with no saturated or light colors are ignored. We plot the results for the average ratings of all themes containing each color, along with the unmixed weights. Note that while the results are noisier, particularly for MTurk due to the fewer constraints, the same relative preference for hues is apparent with more exaggerated peaks and valleys.

2 HSV histograms of Kuler data

In Figures 2 and 3 we plot the distribution of colors with respect to hue versus saturation, and hue versus value for both datasets. The distribution of colors from both datasets is very similar, showing a strong preference for bright warm colors and cyans. Note that fully saturated colors are extremely popular for all hues. However, de-saturated yellows are common, with reds tending to be more saturated. Greens are mostly lighter and unsaturated.

3 Joint hue histograms of Kuler and COLOURLovers data

In Figure 4 we show the joint probability over all hues in a theme. That is, the probability that two hues will be in the same theme, regardless of adjacency. Results are similar to probabilities for adjacent hues with strong diagonal lines present in the Kuler dataset which indicate the use of hue templates (see main text for discussions).

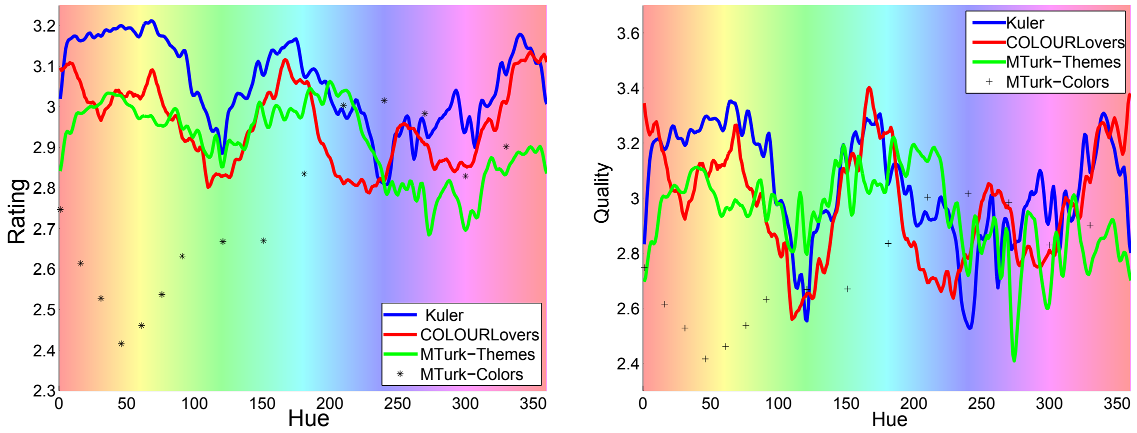

Figure 1: Color preferences. Left: Mean rating of themes containing each hue, and individual color ratings from MTurk. Right: Unmixed rating quality for each hue.

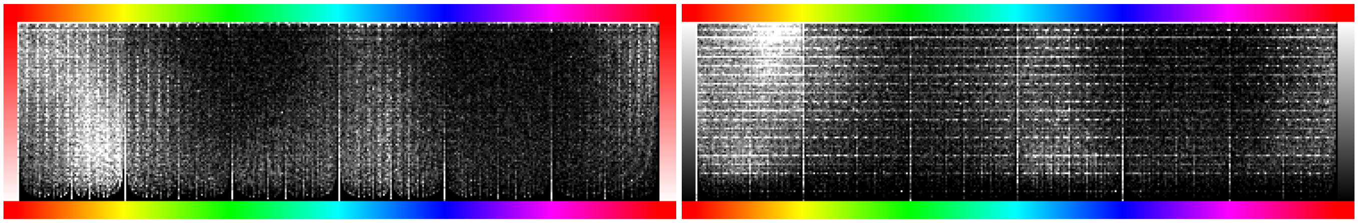

Figure 2: Kuler color density of hue versus saturation (left), hue versus value (right).

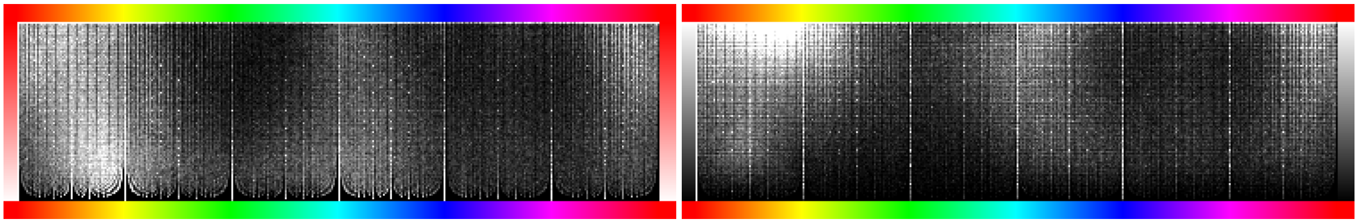

Figure 3: COLOURLovers color density of hue versus saturation (left), hue versus value (right).

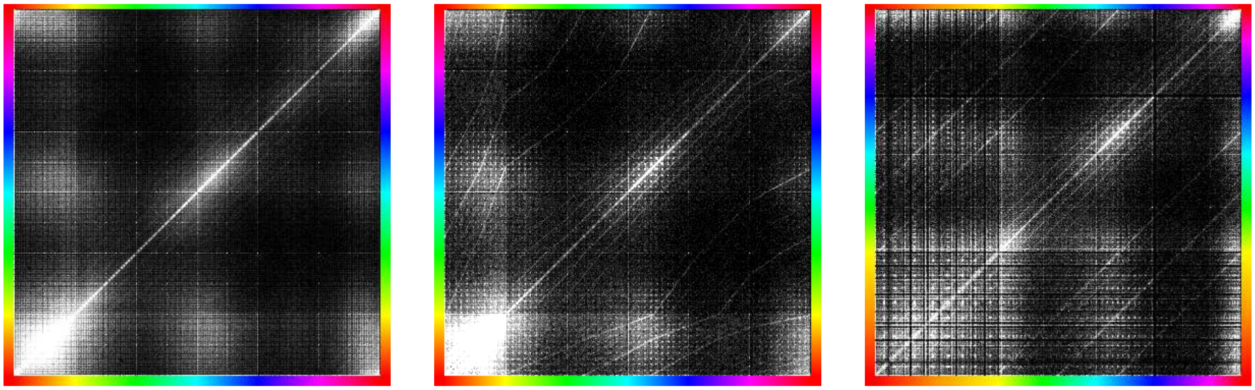

Figure 4: Joint probability over all hues in a theme. Top left, COLOURLovers dataset. Top right, Kuler dataset. Bottom, Kuler dataset with hues remapped to BYR color wheel used in Kuler interface. Diagonal lines indicate hue templates (see main text for discussion)

4 Hue templates

In Figure 5 we show all the hue templates for COLOURLovers, Kuler, and Matsuda. In Figure 6 we show the histogram of template distance for the Kuler and COLOURLovers datasets. Note the spike around zero for templates implemented in the Kuler interface which is mostly lacking in the COLOURLovers data. In the COLOURLovers interface, templates are harder to find and utilize than in Kuler. These results show that people only gravitate towards the most basic templates like i, V, and I, and which are also implemented in both interfaces.

In Figure 7 and 8 we show the breakdown of ratings versus distance for each template. Note that generally, the distance to a template does not appear to be strongly connected to ratings. However, for simple templates like i, V, I which are implemented in Kuler and COLOURLovers, being too close to the template actually results in a lower rating.

We also assign themes to their nearest template and plot the histogram count along with mean ratings with standard deviation and 2 standard error. The results show a great deal of variation but generally, themes distant from a template do not score lower than themes nearer a template. Certain templates are more popular than others, particularly simpler templates like V and L, which both indicate a set of nearby hues. Monochromatic themes (template i) are popular in MTurk, but less popular in COLOURLovers and Kuler. The R and X templates which have 3 and 4 hues spread equally across the hue wheel are among the least popular, as are greyscale themes (template N). We show two thresholds (in Figures 9 and 10. Note that the mean ratings are similar, as are the relative popularity of the templates.

5 Feature weights

See weight.csv in the submitted code and data zip file for weights. The naming convention is to specify the color space first (hsv, chsv, lab, rgb). This is followed by the feature name (for ex, SortedDiff, or StdDev). Next, the dimension of the color space is specified (D1, D2, or D3), followed by the color (C1,C2,C3,C4, or C5) if they are present in the feature. For example, labMedian-D3 indicates the median of the 5 colors of the third dimension in CIELab(B). rgb-D1-C4 indicates the first dimension of RGB space (R) of the fourth color of the theme.

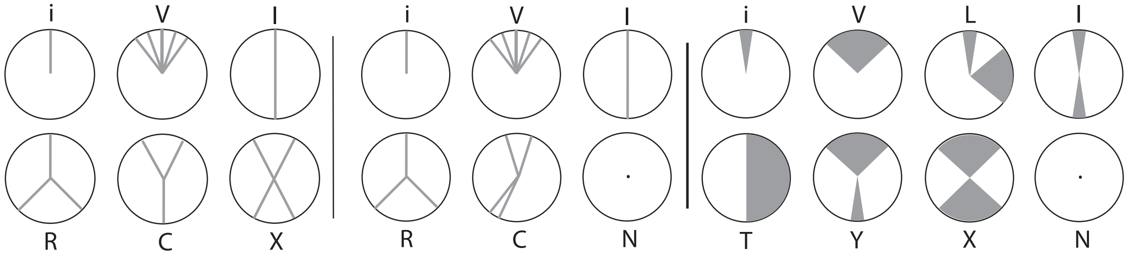

Figure 5: **Hue templates implemented in COLOURLovers(left), Kuler (middle), and those proposed by Matsuda [1995] (right).** Kuler implements several color selection rules (equivalent to Matsuda's i, V, I), as well as others: t(R)iad, (C)ompound. Each theme is described by a color wheel, with gray areas for the hues used by that theme. COLOURLovers implements the i, V, I, R, Y, X templates. Matsuda uses sectors over the hue wheel, whereas Kuler and COLOURLovers use fixed angle distances which matches classical theory. To compare with Matsuda we use the sector centers, or equally spaced hues in the sectors.

6 Minimum Ratings

In Figure 11 we plot the effect of increasing the minimum number of ratings for each theme. A minimum number of 2 ratings was chosen as this provided a large gain over the baseline estimator while still preserving a large number of themes.

7 Color Suggestion Distance

How good are color suggestions made by our model? In the main paper, we show the results of a study applying these to graphic designs. However, another test is to select a random color from a theme, set it to grey, and optimize for the best possible color using our model. Since the themes were human-rated, we have an estimate of the original color's quality. When theme is poorly rated, we expect the original color was badly chosen, so our model will likely choose a more distant color. However, when the theme is highly rated, we expect that the user has chosen a good color. So we expect that on average, our choice would be closer. We can then plot the distance from original to optimized color (in CIELab) compared to the human rating. If the model suggests good colors on average, we expect to see a downward trend.

In Figure 12 we plot the results for themes from the Kuler and MTurk test datasets (4,861 and 4,291 themes respectively). We only use the MTurk and Kuler datasets as both have ground-truth human ratings. Both models have a downward trend which helps validate our model. For Kuler, the increased noise is likely since the low numbers of ratings per theme create more variance along the x-axis.

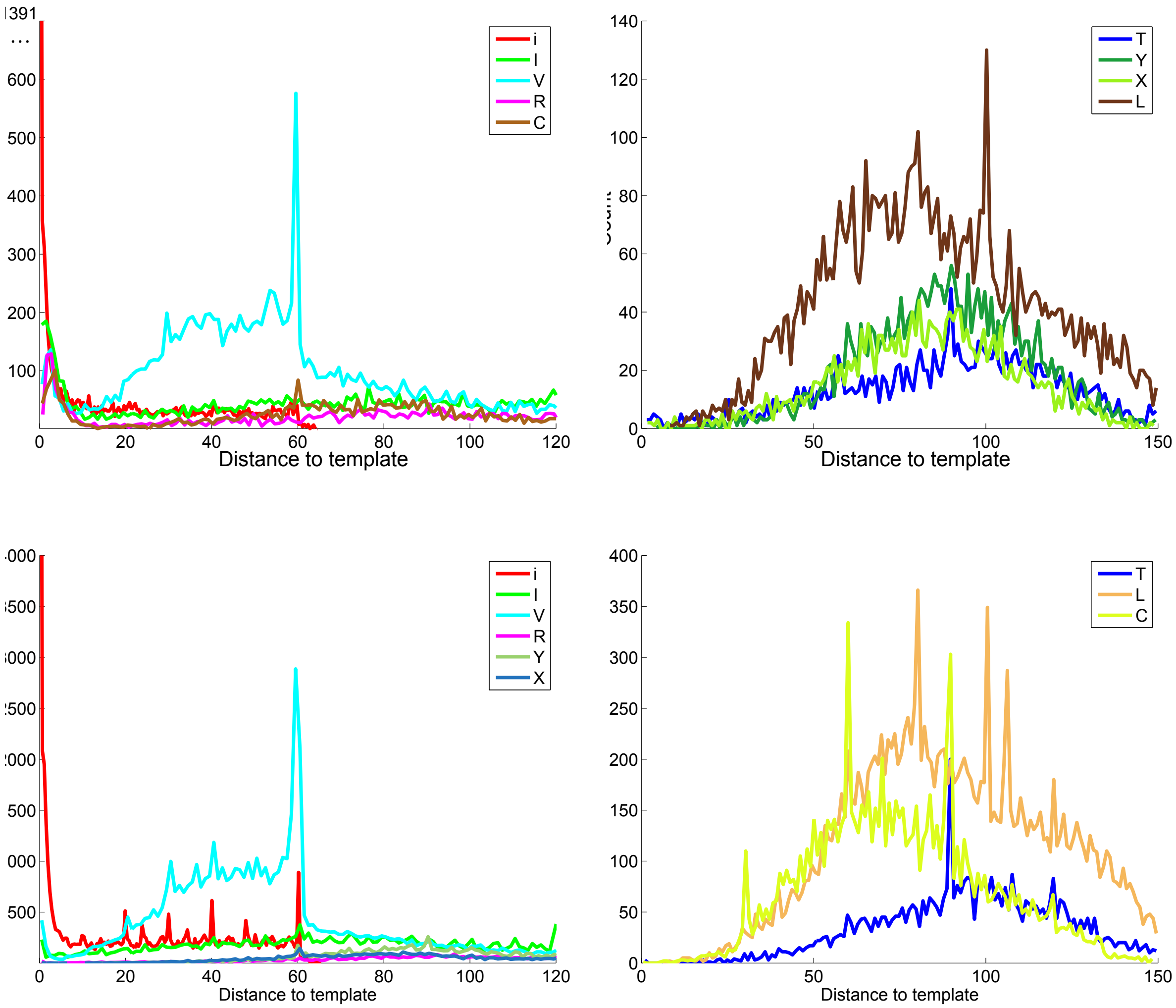

Figure 6: Top row, template distance in Kuler dataset for interface-implemented templates, and for the rest of Matsuda's templates. Bottom row, template distance for COLOURLovers dataset for interface-implemented templates, and for the rest of Matsuda's templates. Note the spike around zero for templates implemented in the Kuler interface which is mostly lacking in the COLOURLovers data.

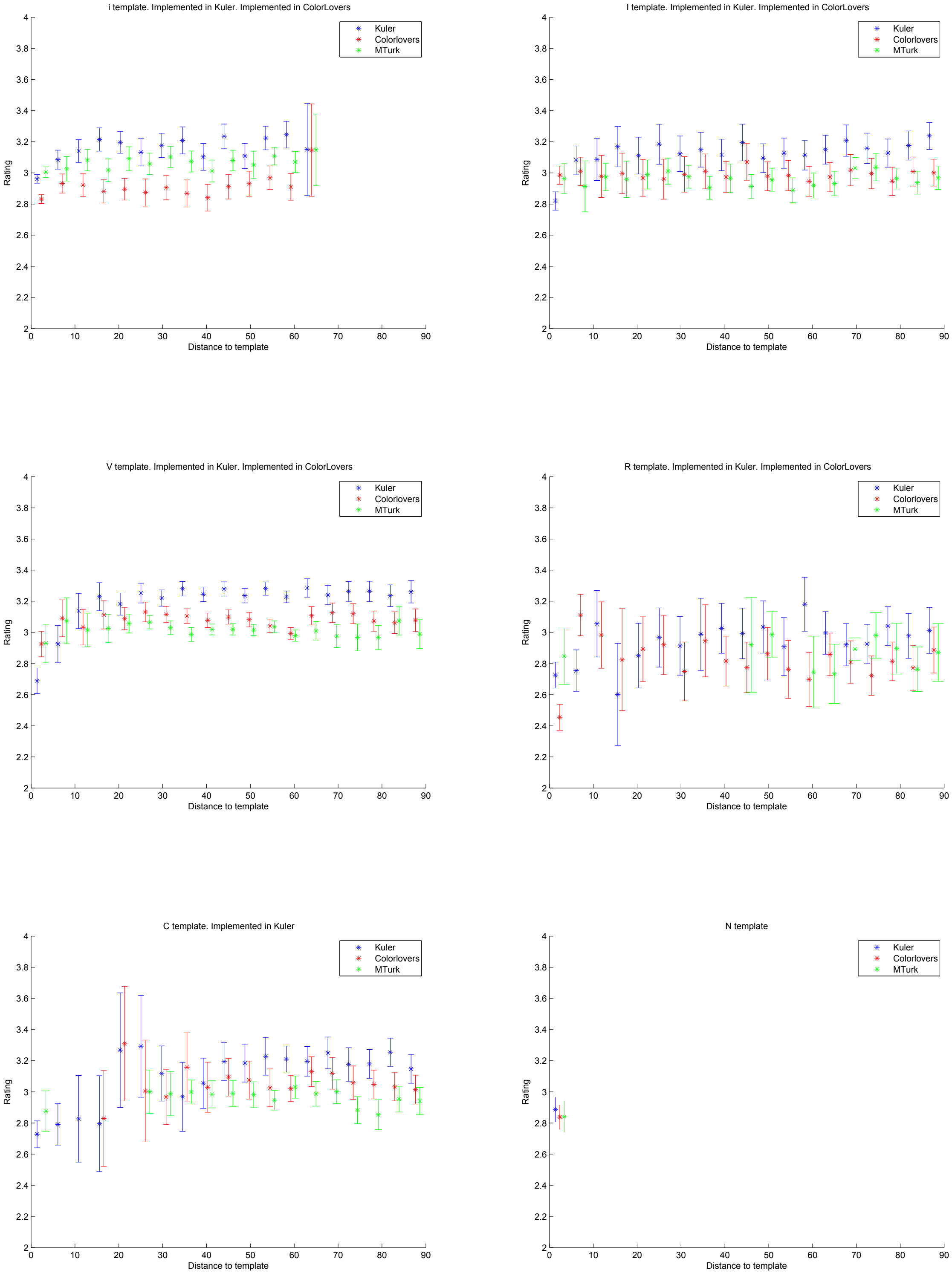

Figure 7: Mean rating versus template distance for each template. Error bars show 2 standard errors.

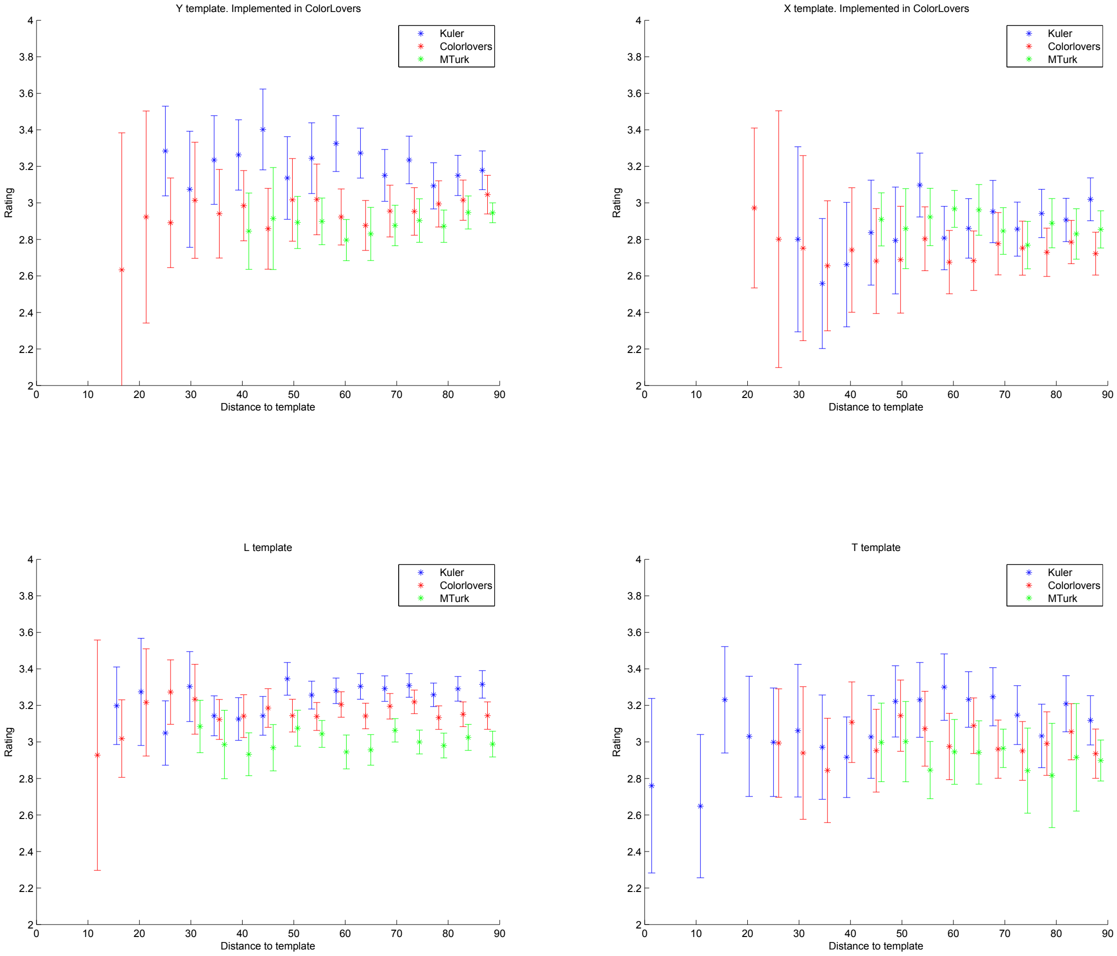

Figure 8: Mean rating versus template distance for each template. Error bars show 2 standard errors.

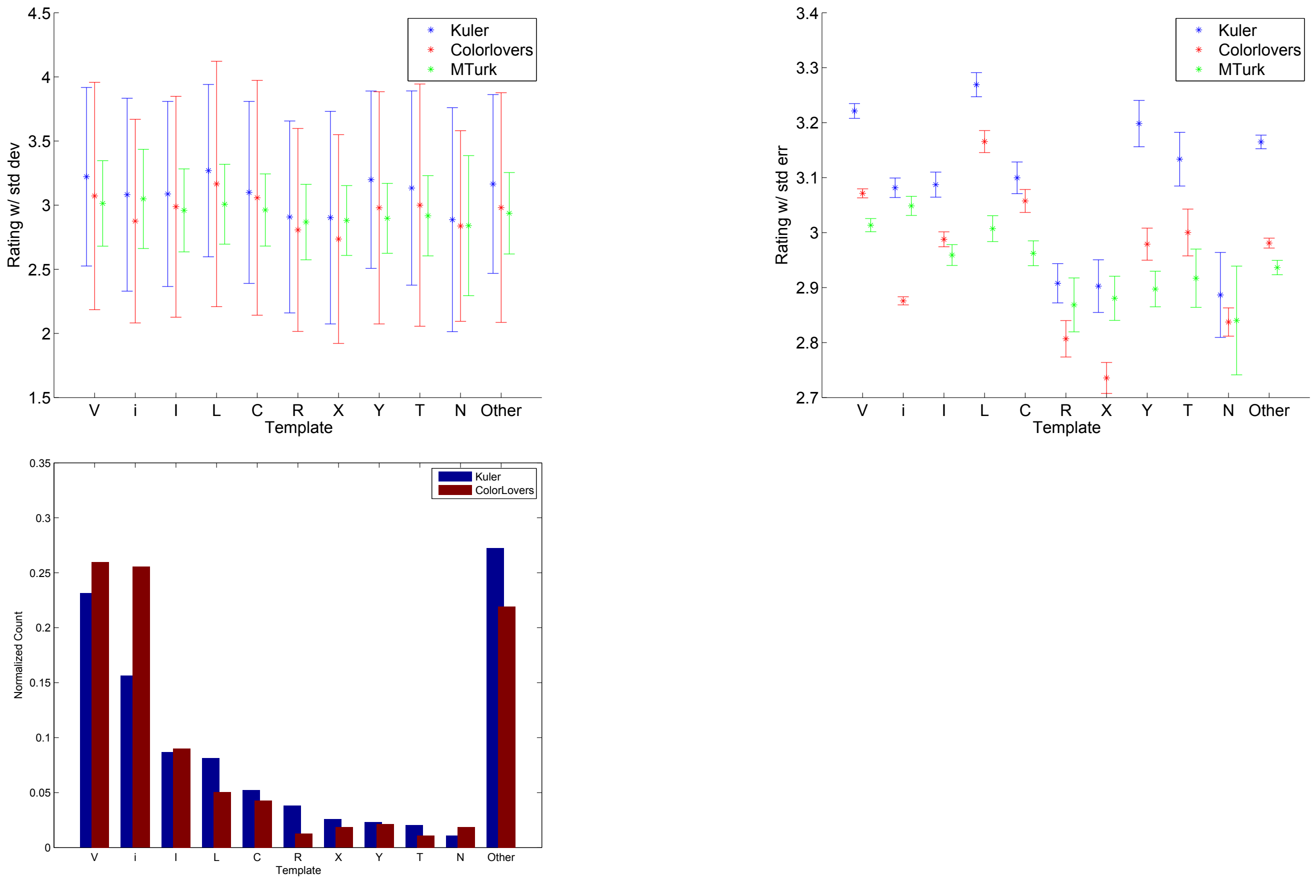

Figure 9: Template mean ratings with standard deviation and 2 standard errors, and histogram count. Themes assigned to template if distance < 90 degrees. See main text for description of distance metric.

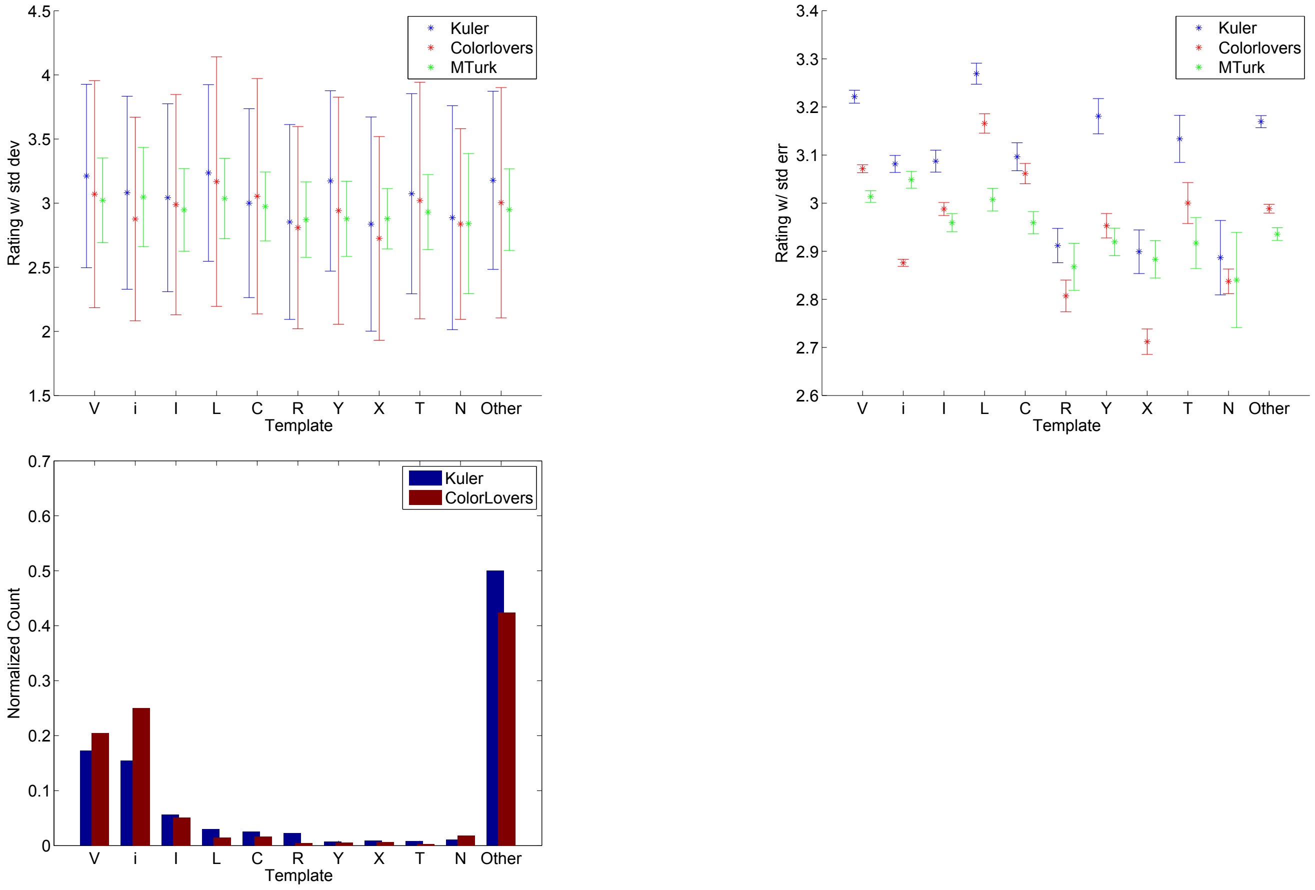

Figure 10: Template mean ratings with standard deviation and 2 standard errors, and histogram count. Themes assigned to template if distance < 60 degrees. See main text for description of distance metric.

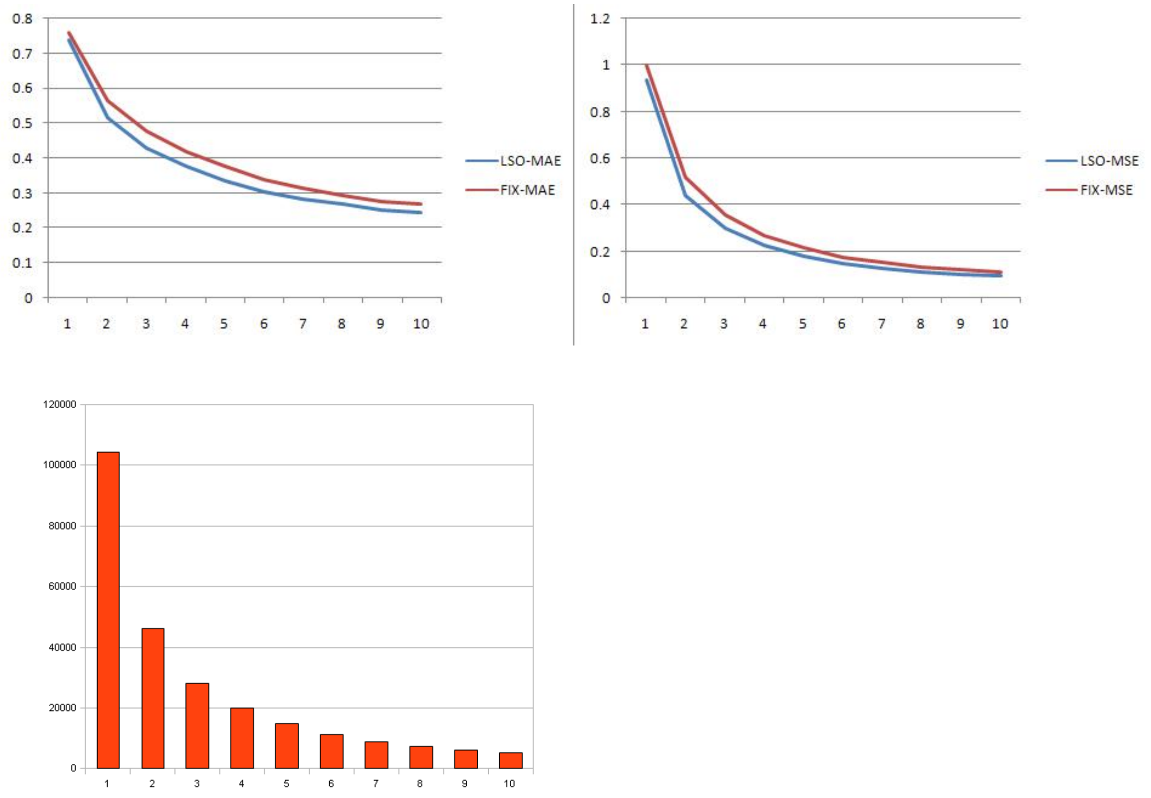

Figure 11: Top, effect of increasing the minimum number of ratings for Kuler dataset. Bottom, histogram of theme count for each test.

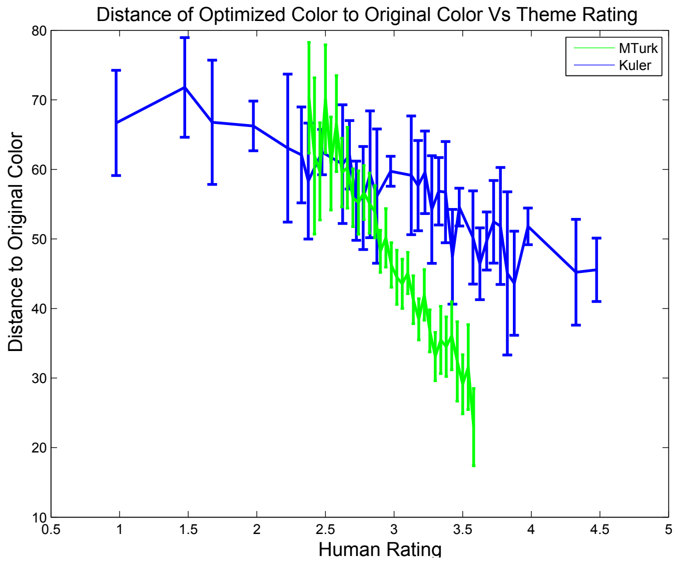

Figure 12: Distance of an optimized color from the original compared to the theme rating. A downward trend indicates that the model generally suggests colors which are closer to the original for highly rated themes (where the original color choice was likely good) than for poorly-rated themes (where the original color choice was likely poor).